

# Draft White Paper 3

## Technological issues for computer-based assessment

Benő Csapó, John Ainley, Randy Bennett, Thibaud Latour, Nancy Law

The Assessment and Teaching of 21st Century Skills project was created by Cisco, Intel and Microsoft and launched at the Learning and Technology World Forum 2009 in London. During 2009, the project operated with five Working Groups, each of which produced a White Paper. These papers will be fully edited into a volume that will be published electronically on the project website ([www.atc21s.org](http://www.atc21s.org)). Print publication is also being considered.

As a report to the Learning and Technology World Forum 2010 in London, final drafts of the papers are collected together in this set and posted on the project website for Forum participants and others who can freely access them on the website. These drafts are not for formal citation. All persons registered on the project website for updates will be advised when the final publication has been posted on the site.

January 2010



# Contents

## Table of Contents

Contents .....	i
Table of Contents .....	i
List of figures .....	ii
Abstract .....	3
Technological issues for computer-based assessment .....	5
Diversity of assessment domains, purposes, and contexts .....	6
Assessment domains .....	6
Assessment purposes .....	8
Assessment contexts .....	8
Using technology to improve assessment .....	8
Formalizing descriptors for technology-based assessment .....	9
Scale .....	10
Theoretical grounds .....	10
Scoring mode .....	10
Reference .....	10
Framework type .....	10
Technology purpose .....	10
Context variables .....	11
Stakeholders .....	11
Intentionality/directionality .....	11
Review of previous research and development .....	11
Research on using technology for assessment .....	12
Assessment of established constructs .....	12
Extending assessment domains .....	13
Assessing new constructs .....	13
Assessing dynamics .....	14
Implementing technology-based assessment .....	14
Technology-based assessments in Australia .....	14
Technology-based assessments in Asia .....	17
Examples of research and development on technology-based assessments in Europe .....	20
Examples of technology in assessment in the US .....	22
Applying technology in international assessment programs .....	23
Technology for item development and test management .....	25
Principles for developing technological platforms .....	25
Enabling assessment of reliability of data and versatility of instruments .....	25
Enabling efficient management of assessment resources .....	25
Accommodating a diversity of assessment situations .....	26
Item building tools .....	26
Balancing usability and flexibility .....	26
Separating item design and item implementation .....	26
Distinguishing authoring from run-time and management platform technologies .....	27
Items as interactive composite hypermedia .....	27
Extending item functionalities with external on-demand services .....	29
Item banks, storing item meta-data .....	31
Delivering technologies .....	33

Factors shaping choice of delivery technology .....	33
Types of delivery technology .....	34
Internet-based delivery .....	34
Local server delivery .....	35
Delivery on removable media .....	35
Provision of mini-labs of computers .....	35
Use of delivery methods.....	36
Task presentation, response capture, and scoring .....	37
Task presentation and response entry.....	37
Domains in which practitioners primarily use specialized tools .....	38
Domains in which technology may be used exclusively or not at all .....	44
Domains in which technology use is central to the definition .....	46
Scoring .....	48
Validity issues raised by the use of technology for assessment.....	49
Special applications and testing situations enabled by new technologies.....	52
Assessing students with special educational needs.....	53
Connecting individuals: assessing collaborative skills and group achievement.....	54
Need for further research and development .....	55
Migration strategies.....	55
Security, availability, accessibility, comparability .....	56
Ensuring framework and instrument compliance with model-driven design.....	57
Potential themes for research projects .....	60
References .....	63

### List of figures

Figure 1: Overview of assessment items for technology literacy.....	18
Figure 2: Overview of grade 5 assessment items for information literacy in mathematics.....	19
Figure 3: Overview of grade 8 assessment items for information literacy in science .....	21
Figure 4: Illustration of eXULIS handling & integrating different media types & services.....	29
Figure 5: Inserting a point on a number line.....	36
Figure 6: A numeric entry task allowing use of an onscreen calculator .....	37
Figure 7: A numeric entry task requiring use of a response template.....	38
Figure 8: Task with numeric entry & many correct answers to be scored automatically .....	40
Figure 9: Task requiring symbolic expression for answer.....	40
Figure 10: Task requiring forced choice and text justification of choice.....	41
Figure 11: Graph construction with mouse clicks to shade/unshade boxes .....	42
Figure 12: Plotting points on grid to create a line or curve.....	43
Figure 13: Item requiring construction of a geometric shape.....	44
Figure 14: A response type for essay writing .....	45
Figure 15: A simulated internet search problem .....	46
Figure 16: Environment for problem solving by conducting simulated experiments.....	47

## Abstract

This paper reviews the contribution of new information-communication technologies to the advancement of educational assessment. Improvements can be described in terms of precision in detecting the actual values of the observed variables, efficiency in collecting and processing information, and speed and frequency of feedback given for the participants and stakeholders. The paper reviews previous research and development in two ways, describing the main tendencies in four continents (Asia, Australia, Europe and the US) and summarizing research on how technology advances assessment in some crucial dimensions (assessment of established constructs, extension of assessment domains, assessment of new constructs and in dynamic situations). As there is a great variety of applications of assessment in education, each one requiring different technological solutions, the paper classifies assessment domains, purposes and contexts and identifies the technological needs and solutions for each. The paper reviews the contribution of technology to the advancement of the entire educational evaluation process from authoring and automatic generation and storing items through delivery methods (Internet-based, local server, removable media, mini-computer labs) and forms of task presentation made possible with technology to response capture, scoring and automated feedback and reporting. The paper also reviews some special cases for which new technologies have enabled significant advances (e.g. assessments of students with special educational needs, assessment of collaborative skills and group achievement) and discusses the validity issues raised by the application of the new technologies (e.g. factors influencing achievements when working with technological tools, the question of transferability of skills measured in a virtual environment). Finally, the paper identifies areas where further research and development is needed (migration strategies, security, availability, accessibility, comparability, framework and instrument compliance) and lists themes for research projects feasible in the *Assessment and Teaching of 21st Century Skills project*.



# Technological issues for computer-based assessment

Benő Csapó, John Ainley, Randy Bennett, Thibaud Latour, Nancy Law

Information-communication technology (ICT) offers so many outstanding possibilities for teaching and learning that its application has been growing steadily in every segment of education. Within the general trends of the utilization of ICT in education, technology-based assessment (TBA) represents a rapidly increasing share. Several traditional assessment processes can be carried out more efficiently by means of computers. In addition, technology offers new assessment methods that cannot be otherwise realized. It is without doubt that TBA will replace paper-based testing in most of the traditional assessment scenarios, and technology will further extend the territories of assessment in education, as it provides frequent and precise feedback for the participants in learning and teaching that cannot be achieved by any other means.

On the other hand, large-scale implementation of TBA still faces several technological challenges that need further research and a lot of experimentation in real educational settings. The basic technological solutions are already available, but their application in everyday educational practice, especially their integration into educationally optimized, consistent systems requires further developmental work.

A variety of technological means appear in schools, and their diversity, compatibility, connectivity and co-working require further considerations. Each new technological innovation finds its way to schools, but not always in a systematic way. Thus, the possibilities of technology-driven modernization of education – when the intent of applying emerging technological tools motivates changes – are limited. In this paper, another approach is taken in which the actual and conceivable future problems of educational development are considered, and the available technological means are evaluated according to their potential to contribute to solving the problems.

Technology may significantly advance educational assessment in a number of dimensions. It improves the precision of detecting the actual values of the observed variables, efficiency of collecting and processing information; it enables the sophisticated analysis of the available data, supports decision making, and provides rapid feedback for the participants and stakeholders. Technology helps to detect and record psychomotor, cognitive and affective characteristics of students and the social contexts of teaching and learning processes alike. When we deal with technological issues of educational assessment, we limit our analysis for the human side of the human-technology interaction. Although technological problems in a narrow sense, like parameters of the available instruments – e.g. processor speed, screen resolution, connection bandwidth – are crucial in educational application, these questions play a secondary role in our study. In this paper we mostly use the more general term *technology based assessment*, meaning that there are several technical tools beyond the most commonly used computers. Nevertheless, we are aware that in the foreseeable future, computers will play a dominant role.

The entire project focuses on the 21st century skills; however, when dealing with technological issues, we have to consider a broader perspective. In this paper, our position concerning the 21st century skills is that we are not dealing exclusively with them, because:

- they are not yet identified with the precision and accuracy that their definition could orient the work concerning technological issues;
- we assume that they are based on certain basic skills and ‘more traditional’ sub-skills, and technology should serve the assessment of such components as well;
- in real educational context, assessment of 21st century skills is not expected to be separated from the assessment of other components of students’ knowledge and skills; therefore, the application of technology should cover a broader spectrum;

- several technologies used for the assessment of students' knowledge today may be developed and adapted for the specific needs of the assessment of 21st century skills; and
- there are skills that are obviously [related](#) to the modern, digital world, and technology offers excellent means to assess them; therefore we deal with these specific issues whenever appropriate throughout the paper (e.g. dynamic problem solving, complex problem solving in technology-rich environment, working in groups where members are connected by ICT).

Different assessment scenarios require different technological conditions, so one single solution cannot optimally serve every possible assessment needs. Teaching and learning in a modern society extend well beyond formal schooling; and even in traditional educational settings, there are diverse forms of assessment, which require technologies adapted to the actual needs. Different technological problems have to be solved when computers are used to administer high-stake, large-scale nationally or regionally representative assessments under standardized conditions, or, low-stake, formative, diagnostic assessment in a classroom environment under diverse school conditions. Therefore, we provide an overview of the most common assessment types and identify their particular technological features.

Innovative assessment instruments raise several methodological questions, and it requires further analysis on how data collection with the new instruments can satisfy the basic assumptions of psychometrics, and on how they fit into the models of classical or modern test theories. This paper, in general, does not deal with methodological questions. There is one methodological issue that should be considered from technological point of view, however, and this is validity. Different validity problems may arise when TBA is applied to replace traditional paper-based assessment and when skills related to the digital world are assessed.

In this paper, technological issues of assessment are considered in a broader sense. Therefore, beyond reviewing the novel data-collection possibilities, we deal with the questions of how technology may serve the entire educational evaluation process, including item generation, automated scoring, data-processing, information flow, feedback, and supporting decision-making.

## **Diversity of assessment domains, purposes, and contexts**

Assessment occurs in diverse domains for a multiplicity of user purposes and in a variety of contexts for those being assessed. Those domains, purposes, and contexts are important to distinguish because they can have implications for how technology might be employed to improve testing and for the issues associated with achieving that improvement.

### **Assessment domains**

The relationship between domain, or construct, definition and technology is critical because that definition influences the role that technology can play in assessment. Below, we distinguish five general situations, each of which poses different implications for the role that technology might play in assessment.

The first situation is characterized by domains in which practitioners interact with new technology primarily through the use of specialized tools, if they use such technology tools at all. In mathematics, such tools as symbol manipulators, graphing calculators, and spreadsheets are frequently used but typically for only certain purposes. For many mathematical problem-solving purposes, paper and pencil remains the most natural and fastest way to address a problem and most students and practitioners use that medium a significant portion of the time. It would be relatively rare for a student to use technology tools exclusively for mathematical problem solving. For domains in this category, testing with technology needs either to be restricted to those problem-solving purposes for which technology is typically used or be implemented in such a way as not to



negatively influence the measurement of those types of problem solving in which technology is not usually employed (Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008).

The second situation is characterized by those domains in which technology may be used exclusively or not at all, depending upon the preferences of the individual. The domain of writing offers the clearest example. Not only do many practitioners and students routinely write on computer, many individuals do virtually their entire academic and workplace writing on computer. Because of the facility provided by the computer, these individuals may write better and faster in that mode than they could on paper. Other individuals still write exclusively on paper. For these students and practitioners, the computer is an impediment because they haven't learned how to use it to compose. For domains of this second category, testing with technology can take three directions, depending upon the information needs of test users: (1) testing all students in the traditional mode to determine how effectively they perform in that mode, (2) testing all students with technology to determine how proficient they are in applying technology in that domain, or (3) testing students in the mode in which they routinely work (Horkay, Bennett, Allen, Kaplan, & Yan, 2006).

The third situation is defined by those domains in which technology is so central to the definition that removing it would render the definition meaningless. The domain of computer programming would be an example. That domain cannot be effectively taught or practiced without using computers. For domains of this category, proficiency cannot be effectively assessed unless all individuals are tested through technology (Bennett, Persky, Weiss, & Jenkins, 2007).

The fourth situation relates to assessing whether someone is capable of achieving a higher level of performance with the appropriate use of general or domain-specific technology tools than would otherwise be possible without them. It differs from the third situation in that the task may be performed without the use of tools, but only by those who have a high level mastery of the domain, and often in rather cumbersome ways. Here the tools are those generally referred to as cognitive tools, such as simulations and modeling tools (Mellar et al., 1994; Feurzeig and Roberts, 1999), geographic information systems (Kerski, 2003; Longley, 2005), visualization tools (Pea, 2002).

The fifth situation relates to the use of technology to support collaboration and knowledge building. It is commonly acknowledged that knowledge creation is a social phenomenon achieved through social interactions, even if no direct collaboration is involved (Popper, 1972). There are various projects on technology-supported learning through collaborative inquiry in which technology plays an important role in the provision of cognitive and metacognitive guidance (e.g. in the WISE project, see Linn and Hsi, 1999). In some cases the technology plays a pivotal role in supporting the socio-metacognitive dynamics that are found to be critical to productive knowledge building (Scardamalia & Bereiter, 2003), since knowledge building is not something that happens naturally, but rather, has to be an intentional activity at the community level (Scardamalia, 2002).

Thus, how a domain is practiced, taught, and learned influences how it should be assessed because misalignment of assessment and practice methods and can compromise the meaning of assessment results. Also, it is important to note that over time, domain definitions change because how a domain is practiced and taught changes, a result in part of the emergence of new technology tools suited to the domain. Domains that today are characterized by the use of technology for only specialized purposes may tomorrow see a significant proportion of individuals employing technology as their only means of practice. As tools advance, technology could become central to the definition of that domain too .

Of the five domains of technology use described above, the third, fourth and fifth domains pose the greatest challenge to assessment, and yet it is exactly these domains of technology use that are most important to include in the assessment of 21st century skills, since "the real promise of technology in education lies in its potential to facilitate fundamental, qualitative changes in the nature of teaching and learning" (Panel on Educational Technology of the President's Committee of Advisors on Science and Technology, 1997, p.33).

## Assessment purposes

Here, we distinguish four general purposes for assessment deriving from the two-way classification of assessment “object” and assessment “type.” The object of assessment may be the student or it may be a program or institution. Tests administered for purposes of drawing conclusions about programs or institutions have traditionally been termed “program evaluation.” Tests given for drawing conclusions about individuals have often been called “assessment.”

Within each of program evaluation and assessment, two types can be identified: formative versus summative (Bloom, 1969; Scriven, 1967). Formative evaluation centers upon providing information for purposes of program improvement, whereas summative evaluation focus on judging the overall value of a program. Similarly, formative assessment is intended to provide information of use to the teacher or student in modifying instruction, whereas summative assessment centers upon documenting what a student (or group of students) knows and can do.

## Assessment contexts

Assessment context generally refers to the stakes associated with the decisions that are based on test performance. The highest stakes are associated with those decisions that are serious in terms of their impact on individuals, programs or institutions and that are not easily reversible. The lowest stakes are connected to decisions that are likely to have less impact and that are easily reversible. While summative types have typically been taken as high stakes and formative types as low stakes, such blanket classifications may not always hold, if only because a single test may have different stakes for different constituencies. The US National Assessment of Educational Progress (NAEP) is one example of a summative test in which performance is of low stakes to students, as no individual scores are computed, but high stakes for policy makers whose states are publicly ranked. A similar situation obtains for summative tests administered under the US *No Child Left Behind* act, where the results may carry no consequence for students but major consequences for individual teachers, administrators, and schools. On the other hand, a formative assessment may itself take on considerable stakes for a student (but low stakes for the school) if the assessment directs that student toward developing one skill to the neglect of another skill more critical to that student’s near-term success (e.g., in preparing for an upcoming musical audition).

The above definition of context is possibly adequate if the assessment domain is well understood and assessment methods are well developed. If the domains of assessment and/or assessment methods (such as using digital technology to mediate the delivery of the assessment) are new, however, rather different considerations in the design and method are called for. To measure more complex understanding and skills and to integrate the use of technology into the assessment process to reflect such new learning outcomes requires innovation in assessment (Quellmalz & Haertel, 2004). In such situations, assessment instruments probably have to be developed or invented, and it is obvious that both the validity and reliability can only be refined and established over a period of time, even if the new assessment domain is well defined. In the case of assessing 21st century skills, this kind of contextual challenge is even greater since what constitutes the skills to be assessed are in themselves a subject of debate. How innovative assessment can provide formative feedback on curriculum innovation and vice versa is another related challenge.

## Using technology to improve assessment

Technology can be used to improve assessment in at least two major ways: by changing the business of assessment and by changing the substance of assessment itself (Bennett, 2001). The business of assessment refers to the core processes that define the enterprise. Technology can help make these core processes more efficient. Examples can be found in:

- developing tests, making the questions easier to generate automatically or semi-automatically, share, review, and revise (e.g., Bejar, Morley, Wagner, Bennett, & Revuelta, 2003);

- delivering tests, obviating the need for printing, warehousing, and shipping paper;
- presenting dynamic stimuli like audio, video, and animation, making obsolete the need for the specialized equipment currently used in some testing programs that assess such constructs as speech and listening (e.g., audio cassette recorders, VCRs) (Bennett, Goodman, Hessinger, Liggett, Marshall, Kahn, & Zack, 1999);
- scoring constructed responses on screen, allowing marking quality to be monitored in real time and potentially eliminating the need to gather examiners together (Zhang, Powers, Wright, & Morgan, 2003);
- scoring some types of constructed responses automatically, reducing the need for human reading (Williamson, Mislevy, & Bejar, 2006); and
- distributing test results, cutting the costs of printing and mailing reports .

Changing the substance of assessment involves using technology to change the nature of what is tested, or learned, in ways not practical with traditional assessment approaches or with technology-based duplications of those approaches (e.g., using a computer to record an examinee's speech in the same way as a tape recorder is now used). An example would be asking students to experiment with and draw conclusions from an interactive simulation of a scientific phenomenon they could otherwise not experience, and then using features of their problem-solving processes in making judgments about those students (e.g., Bennett, Persky, Weiss, & Jenkins, 2007). A second example would be in structuring the test design so that, by virtue of the way in which the assessment responds to student actions, students learn in the process of taking the assessment.

The use of technology in assessment may also play a crucial role in informing curriculum reform and pedagogical innovation, particularly in areas in which technology becomes crucial to the learning in specific domains. For example, the Hong Kong SAR government commissioned a study to conduct online performance assessment of students' information literacy skills as part of the evaluation of the effectiveness of its IT in education strategies (Law et al., 2007). In Hong Kong, an important premise for the massive investments to integrate IT in teaching and learning is to foster the development of information literacy skills in students so that they can become more effective lifelong learners and that they can accomplish the learning in the designated curriculum more effectively. The study assessed students' ability to search for and evaluate information, and to communicate and collaborate with distributed peers in the context of authentic problem solving through an online platform. The study found that while a large majority of the assessed students were able to demonstrate basic technical operational skills, their ability to demonstrate higher levels of cognitive functioning such as evaluation and integration of information was rather weak. This led to new initiatives in the Third IT in Education Strategy (EDB, 2008) to develop curriculum resources and self-access assessment tools on information literacy. This is an example in which assessment is used formatively to inform and improve on education policy initiatives.

How technology might be used to improve assessment, and address the issues encountered in so doing, interacts with the domain, the purpose, and the context for assessment. For example, fewer issues might be encountered upon implementing formative assessments in low stakes contexts targeted at domains where technology is central to the domain definition as compared with summative assessments in high stakes contexts where technology typically is used only for certain types of problem solving.

### **Formalizing descriptors for technology-based assessment**

Assessment in general and computer-based assessment in particular are characterized by a large number of variables that influence decisions on many aspects of organization, methodology and technology. In turn these decisions strongly influence the level of risk and risk management, change management, costs and timelines. Decisions on the global design of an evaluation program can be considered as a bijection between the assessment characteristic space and the assessment design space. ( $d=C \otimes D$ ,  $D=\{O,M,T\}$ ) In order to scope and address assessment challenges and support decision-making better, beyond the inherent characteristics of the framework and instrument

themselves, one needs to define a series of dimensions describing the space of assessment. It is not the purpose of this paper to discuss thoroughly each of these dimensions and their relationship with technologies, methods, instruments and organizational processes. It is important, however, to describe briefly the most important aspects of assessment descriptors. A more detailed and integrated analysis should be performed to establish best practice recommendations. In addition to the above-mentioned descriptors, one can also cite the following ones.

### *Scale*

The scale of an assessment should not be confused with its objective. Indeed, when considering assessment objectives, one considers the level of *granularity* of the relevant and meaningful information that is collected and analyzed during the evaluation. Depending on the assessment object, the lowest level of granularity, i.e. the elementary piece of information, may either be individual scores or average scores over populations or sub-populations, considered as systems or sub-systems. The scale of the assessment depicts the *number* of information units collected, somehow related to the size of the sample. Exams at school levels and certification are typically small-scale assessments, while PISA or NAEP are typically large-scale operations.

### *Theoretical grounds*

This assessment descriptor corresponds to the theoretical framework used to set up the measurement scale. *Classical* assessment uses a possibly weighted ratio of correct answers vs. total number of questions while Item Response Theory (IRT) uses statistical parameterization of items. As a sub-descriptor, scoring method must be considered from theoretical and procedural or algorithmic points of view.

### *Scoring mode*

Scoring of the items and of the entire test, in addition to reference models and procedures, can be *automatic*, *semi-automatic*, or *manual*. Depending on this organizational processes and technological support, as well as risks on security and measurement quality may change dramatically.

### *Reference*

In some situations, the data collected does not constitute objective evidence of the achievement on the scale or metrics. Subjective evaluations are based on a test-taker assertion about his own level of achievement or potentially about another's level of achievement in the case of hetero-evaluation. These situations refer to declarative assessment, while scores inferred from facts and observation collected by another agent than the test-taker himself are referred to as evidence-based assessments.

### *Framework type*

Assessments are designed in different contexts and for different purposes based on a reference description of the competency, skill, or ability that one intends to measure. These various frameworks have different origins, among which the most important ones are *educational programs and training specifications* (content-based or goal-oriented); *cognitive constructs*; and *skill cards and job descriptions*. The framework type, may have strong implications for organizational processes, methodology and technical aspects of the instruments.

### *Technology purpose*

The place of technology in assessment operations is another very important factor that has an impact on the organizational, methodological and technological aspects of the assessment. While

many variations can be observed, two typical situations can be identified: *computer-aided assessment* and *computer-based assessment*. In the former, the technology is essentially used at the level of organization and operational support processes. The assessment instrument remains paper-and-pencil and IT is only used as a support tool for the survey. In the latter situation, the computer is used to deliver the instrument itself.

### *Context variables*

Depending on the scale of the survey, a series of scaling variables related to the context are also of great importance. Typical variables of this type are *multilingualism*; *multi-cultural aspects*; consideration of *disabilities*; *geographical aspects* (remoteness); *geopolitical, political and legal aspects*; *data collection mode* (e.g. centralized, network-based, in-house).

### *Stakeholders*

The identification of the stakeholders and their characteristics is important for organizational, methodological and technological aspects. Typical stakeholders are the *test taker*, the *test administrator*, and the *test backer*.

### *Intentionality/directionality*

Depending on the roles and relationships between stakeholders, the assessment will imply different intentions and different risks that have to be managed. Typical situations can be depicted using two fundamental questions: (a) which stakeholder assigns the assessment to which stakeholder? (b) which stakeholder evaluates which stakeholder (in other words, which stakeholder provides the evidence or data that are collected during the assessment about which stakeholder)? As an illustration this yields the definition of *self-assessment* where the test taker assigns a test to himself (be it declarative or evidence-based) and manipulates the instrument; or *hetero-assessment* (most generally declarative) where the respondent provides information to evaluate somebody else. In most classical situations, the test taker is different from the stakeholder who assigns the test.

## **Review of previous research and development**

Research and development is reviewed here from two different aspects. On the one hand, a large number of research projects have been dealing with the application of technology for assessment. The devices applied in the experiments may range from the most common, broadly available computers to the emerging cutting edge technologies. In the research context, newly developed expensive instruments may be used and specially trained teachers may participate; therefore these experiments are often small scale, carried out in laboratory context, or may involve only a few classes or schools.

On the other hand, there are efforts for system-wide implementation of TBA, either to extend, improve or replace the already existing assessment systems, or to create entirely new assessment systems. These implementation processes usually involve nationally representative samples of up to a thousand or several thousands of students. Large international programs aim at using technologies for assessment as well, both with the intention of replacing paper-based assessment by TBA and for introducing innovative domains and contexts that cannot be assessed by traditional testing methods. In large-scale implementation efforts, the general educational contexts (school infrastructure) are usually given, and either the existing equipment are used as they are, or new equipment is installed for assessment purposes. Logistics in these cases plays a crucial role; furthermore, several financial and organizational aspects that influence the choice of the applicable technology have to be considered.

## Research on using technology for assessment

ICT has already begun to alter, and has potential to change further, educational assessment. One aspect of this change has been the more effective and efficient delivery of traditional assessments (Bridgeman, 2009). A second aspect has been the use of ICT to expand and enrich assessment tools so that assessments reflect better the intended domains and include more authentic tasks (Pellegrino, Chudowosky & Glaser, 2004). A third aspect has been the assessment of constructs that have either been difficult to assess or which have emerged as part of the information age. A fourth aspect has been the use of ICT to investigate the dynamic interactions between student and assessment material.

Published research literature on technology and computer-based assessment currently reflects a predominance of research comparing the results of paper-based and computer-based assessment of the same construct. This literature seeks to identify the extent to which these two broad modalities provide congruent measures. Some of that literature draws attention to the importance of technological issues (within computer based assessments) on measurement. There is somewhat less literature concerned with the properties of assessments that deliberately seek to extend the construct being assessed by making use of the possibilities that arise from computer-based assessment. An even more recent development has been the use of computer-based methods to assess new constructs: those linked to information technology, those using computer-based methods to assess constructs that have been previously hard to measure or those based on the analysis of dynamic interactions. The research literature on these developments is limited at this stage but will grow as the applications grow.

### *Assessment of established constructs*

One important topic in the efficient delivery of assessments has been that of “equivalence” or whether the scores on computer-administered assessments are comparable with scores on the corresponding paper-based tests. The conclusion of two meta-analyses of studies of computer-based assessments of reading and mathematics among school students is that overall the mode of delivery does not impact on scores greatly (Wang, Jiao, Young, Brooks, & Olson, 2007; Wang, Jiao, Young, Brooks, & Olson, 2008). This generalization appears to hold for small-scale studies of abilities (Singleton, 2001), large-scale assessments of abilities (Csapó, Molnár & R. Tóth, 2009) and large-scale assessments of achievement (Poggio, Glasnapp, Yang, & Poggio, 2004). The same generalization appears to have been found in studies conducted in higher education (Putchinski, Martin, & Moskal, 2007). Despite this overall result there do appear to be some differences in scores associated with some types of questions and some aspects of the way students approach tasks (Johnson & Green, 2007). In particular there appears to be an effect of computer familiarity on performance in writing tasks (Horkay, Bennett, Allen, Kaplan, & Yan, 2006) .

Computer-based assessment, in combination with modern measurement theory, has given impetus to expanding the possibility of computer adaptive testing (Wainer, 2000; Eggen & Straetmans, 2009). In computer adaptive testing student performance on items is used dynamically so that subsequent items are selected from an item bank at a difficulty appropriate for the student thus providing more time-efficient and accurate assessments of proficiency. Adaptive tests can provide more evenly spread precision across the performance range, are shorter for each person assessed maintain a higher level of precision overall than a fixed-form test (Weiss & Kingsbury, 2004). However, they are dependent on building and calibrating an extensive item bank.

There have been a number of studies of variations within a given overall delivery mode that impact on a student’s experience of an assessment. There is wide acceptance of the imperative that all students should experience the tasks or items presented in a computer-based assessment in an identical manner. Uniformity of presentation is assured when students are presented with assessment tasks or items in a test booklet. However, there is some evidence that computer based assessment can impact upon student performance because of variations in presentation that are not relevant to the construct being assessed (Bridgeman, Lennon and Jackenthal 2003; McDonald

2002). Bridgeman and colleagues (2003) point to the influence of variations in screen size, screen resolution and display rate on performance on computer-based assessments. These are issues that arise in computer-based assessments that do not normally arise in pen and paper assessments. Thompson and Weiss (2009) argue that the possibilities of variations in the assessment experience are especially an issue for internet or web-based delivery of assessments. These are important considerations for the design of assessment delivery systems. Large-scale assessments using ICT face the issue of providing a uniform testing environment when school computing facilities may vary considerably.

### *Extending assessment domains*

One of the issues confronting assessment has been that what has been able to be assessed by paper-based methods represents a narrower conception of the domain than one would ideally wish to assess. The practice of assessment has been limited by what could be presented in a printed form and what could be answered by students in a written form. Attempts to provide assessments of broader aspects of expertise have been limited by the need to be consistent and, in the case of large-scale studies, the capacity to process rich answers. In many cases these pressures have resulted in the use of closed-response formats (such as multiple choice) rather than constructed response formats (where students write a short or an extended answer).

ICT can be used to present richer stimulus material (e.g. video or richer graphics), to provide for students to interact with the assessment material and to develop products that are saved for subsequent assessment by raters. In PISA 2006 a computer-based assessment of science (CBAS) was developed for and applied in a field trial in 13 countries. It was then adopted as part of the main study in three countries. CBAS was intended to assess aspects of science that could not be assessed in paper-based formats. It therefore involved an extension of the implemented assessment domain but did not attempt to cover the whole of the intended domain. It was based on providing rich stimulus material linked to conventional test item formats. The design for the field trial included a rotated design that had half of the students doing a paper-based test first, followed by a computer test, and half doing the tests in the opposite order. In the field trial the correlation between the paper-based and computer-based items was 0.90 but it was also found that a two dimensional model (dimensions corresponding to the paper and computer-based assessment items) was a better fit than a one-dimensional model (Martin, 2009). This suggests that the dimension of science knowledge and understanding represented in the CBAS items was related to, but somewhat different from, the dimension represented in the paper-based items. Halldórsson, McKelvie and Björnsson (2009) showed that in the main PISA survey in Iceland boys performed relatively better than girls but this difference was not associated with differences in computer familiarity, motivation or effort. It did appear to be associated with the lower reading load on the computer-based assessment. In other words the difference was not a result of the mode of delivery as such but of a feature that was associated with the delivery mode: the amount of text to be read. At present, reading is modified on computer because of restrictions of screen size and the need to scroll to see what would be directly visible in a paper form. The restriction in the electronic form is likely to be removed as *e-book* and other developments are advanced.

### *Assessing new constructs*

A third focus on research on computer-based assessment is on assessing new constructs. Some of these relate directly to skills either associated with information technology or which have changed in nature as a result of the introduction of information technology. An example of such a construct is "problem solving in rich technology environments" (Bennett, Persky, Weiss & Jenkins, 2007). Bennett and colleagues (2007) measured this construct in a nationally (USA) representative sample of Grade 8 students. The assessment was based on two extended scenarios set in the context of scientific investigation; one involving a search and the other a simulation. The OECD *Programme for International Assessment of Adult Competencies* (PIAAC) includes "problem solving in technology-rich environments" as one of the capabilities that it assesses among adults (OECD, 2008). This refers to the cognitive skills required in the information age and has a focus on solving

problems using multiple sources of information on a laptop computer. The problems are intended to involve accessing, evaluating, retrieving and processing information and incorporate technological and cognitive demands .

Wirth and Klieme (2003) investigated analytical and dynamic aspects of problem solving. Analytical problem solving abilities were those needed to structure, represent and integrate information whereas dynamic problem solving involved the ability to adapt the problem solving process to a changing environment by processing feedback information (and included aspects of self-regulated learning). As a German national option of PISA 2000, the analytical and dynamic problem solving competencies of 15-year-old students were tested using paper-and-pencil tests as well as computer-based assessments. Wirth and Klieme reported that analytical aspects of problem solving competence were strongly correlated with reasoning, while dynamic problem solving reflected a dimension of self-regulated exploration and control that could be identified in computer-simulated domains.

Another example of computer-based assessment involves using new technology to assess more enduring constructs such as teamwork (Kyllonen, 2009). *Situational Judgment Tests* (SJTs) involve presenting a scenario (incorporating audio or video) that involves a problem and asking the student the best way to solve the problem. A meta-analysis of the results of several studies of SJTs of teamwork and concluded that they involve both cognitive ability and personality attributes and that they predict real world outcomes (McDaniel, Hartman, Whetzel & Grubb, 2007). Kyllonen argues that SJTs provide a powerful basis for measuring other constructs such as creativity, communication and leadership provided that it is possible to identify critical incidents that relate to the construct being assessed (Kyllonen & Lee, 2005) .

### *Assessing dynamics*

A fourth aspect of computer based assessment is the possibility of assessing more than an answer or a product but to use information about the process involved to provide an assessment. This information is based on the analysis of times and sequences in data records in logs that track students' paths through a task, choices of which material to access, and decisions about when to start writing an answer (M. Ainley, 2006; Hadwin, Wynne & Nesbitt, 2005). M. Ainley draws attention to two issues associated with the use of time trace data: the reliability and validity of single item measures (which is necessarily the basis of trace records) and appropriate analytic methods for data that span a whole task and use the trend, continuities, discontinuities and contingencies in those data. Kyllonen (2009) identifies two other approaches to assessment that make use of time records available from computer based assessments. One is through the study of the times taken to complete tasks. The other is by using the time to choose between pairs of options to provide an assessment of attitudes or preferences as in the *Implicit Association Test* (IAT).

## **Implementing technology-based assessment**

### *Technology-based assessments in Australia*

Australian education systems have placed considerable emphasis on the application of ICT in education in successive iterations of the *National Goals for Schooling* (MCEETYA, 1999; MCEECDYA, 2008). The national goals adopted in 1999 stated that when students leave school they should 'be confident, creative and productive users of new technologies, particularly information and communication technologies, and understand the impact of those technologies on society' (MCEETYA 1999). This was re-iterated in the more recent *Declaration on Educational Goals for Young Australians* which asserted that "in this digital age young people need to be highly skilled in the use of ICT" (MCEECDYA, 2008).

The implementation of ICT in education was guided by a plan entitled *Learning in an On-line World* (MCEETYA, 2000; 2005) and supported by the establishment of a national company (*education.au*)



to operate a resource network (*Education Network Australia* or *EdNA*) and venture called the *Learning Federation* to develop digital learning objects for use in schools. More recently a *Digital Revolution* has been included as a feature of *National Education Reform Agenda* which is adding impetus to the use of ICT in education through support for improving ICT resources in schools, enhanced internet connectivity and building programs of teacher professional learning. Part of the context for these developments is the extent to which young people in Australia have access to and use ICT (and web-based technology in particular) at home and at school. Australian teenagers continue to have access to, and use, ICT to a greater extent than their peers in most other countries and are among the highest users of ICT in the OECD (Anderson & Ainley, 2009). It is also evident that Australian teachers (at least teachers of mathematics and science in lower secondary school) are among the highest users of ICT in teaching (Ainley, Eveleigh, Freeman & O'Malley, 2009).

In 2005, Australia began a cycle of three-yearly national surveys of the ICT literacy of students (MCEETYA, 2007). Prior to the 2005 national assessment the Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA) defined as ICT as technologies used for accessing, gathering, manipulation and presentation or communication of information and adopted a definition of ICT Literacy as: *the ability of individuals to use ICT appropriately to access, manage, integrate and evaluate information, develop new understandings, and communicate with others in order to participate effectively in society* (MCEETYA, 2007). This definition, which draws heavily on the Framework for ICT Literacy developed by the International ICT Literacy Panel and the OECD PISA ICT Literacy Feasibility Study (International ICT Literacy Panel, 2002). ICT literacy is increasingly regarded as a broad set of generalisable and transferable knowledge, skills and understandings that are used to manage and communicate the cross-disciplinary commodity that is information. The integration of information and process is seen to transcend the application of ICT within any single learning discipline (Markauskaite, 2007). Common to information literacy are the processes of identifying information needs, searching for and locating information and evaluating the quality of information as well transforming information and using it to communicate ideas (Catts and Lau 2008). According to Catts and Lau (2008) "people can be information literate in the absence of ICT, but the volume and variable quality of digital information, and its role in knowledge societies, has highlighted the need for all people to achieve information literacy skills".

The Australian assessment framework envisaged ICT literacy as comprising six key processes: accessing information (identifying information requirements and knowing how to find and retrieve information); managing information (organizing and storing information for retrieval and reuse); evaluating (reflecting on the processes used to design and construct ICT solutions and judgments regarding the integrity, relevance and usefulness of information); developing new understandings (creating information and knowledge by synthesizing, adapting, applying, designing, inventing or authoring); communicating (exchanging information by sharing knowledge and creating information products to suit the audience, the context and the medium); and using ICT appropriately (critical, reflective and strategic ICT decisions and considering social, legal and ethical issues). Progress was envisaged in terms of levels of increasing complexity and sophistication in three strands of ICT use: (a) working with information; (b) creating and sharing information; and (c) using ICT responsibly. In *Working with Information*, students progress from using key words to retrieve information from a specified source, through identifying search question terms and suitable sources, to using a range of specialized sourcing tools and seeking confirmation of the credibility of information from external sources. In *Creating and Sharing Information*, students progress from using functions within software to edit, format, adapt and generate work for a specific purpose, through integrating and interpreting information from multiple sources with the selection and combination of software and tools, to using specialized tools to control, expand and author information, producing representations of complex phenomena. In *Using ICT Responsibly*, students' progress from understanding and using basic terminology and uses of ICT in everyday life, through recognizing responsible use of ICT in particular contexts, to understanding the impact and influence of ICT over time and the social, economic and ethical issues associated with its use. These results can inform the refinement of a development progression of the type discussed in White Paper 2.

In the assessment students completed all tasks on computer using a seamless combination of simulated and live software applications<sup>1</sup>. The tasks were grouped in thematically linked modules each of which followed a linear narrative sequence. The narrative sequence in each module typically involved students collecting and appraising information before synthesizing and reframing the information to suit a particular communicative purpose and given software genre. The overarching narratives across the modules covered a range of school-based and out-of-school based themes. The assessment included items (such simulated software operations) that were automatically scored and items that required constructed responses stored as text or as authentic software artifacts. The constructed response text and artifacts were marked by human assessors .

All students first completed a General Skills Test and then two randomly assigned (Grade appropriate) thematic modules. One reason for conducting the assessment with a number of modules was to ensure that the assessment instrument accessed what was common to the ICT Literacy construct across a sufficient breadth of contexts.

The modules followed a basic structure in which the simulation, multiple-choice and short-constructed response items led up to a single large task using at least one live software application. Typically the lead-up tasks required students to: manage files; perform simple software functions (such as inserting pictures into files); search for information; collect and collate information; evaluate and analyze information; and perform some simple reshaping of information (such as drawing a chart to represent numerical data). The large tasks that provided the global purpose of the modules were then completed using live software. When completing the large tasks, students typically needed to select, assimilate and synthesize the information they had been working with in the lead-up tasks and reframe the information to fulfill a specified communicative purpose. Students spent between 40 per cent and 50 per cent of the time allocated for the module on the large task. The modules with the associated tasks were:

- Flag Design (Grade 6). Students use purpose-built previously unseen flag design graphics software to create a flag.
- Photo Album (Grade 6 & 10). Students use unseen photo album software to create a photo album to convince their cousin to come on holiday with them.
- DVD Day (Grade 6 & 10). Students navigate a closed web environment to find information and complete a report template.
- Conservation Project (Grade 6 & 10). Students navigate a closed web environment and use information provided in a spreadsheet to complete a report to the Principal using Word.
- Video Games and Violence (Grade 10). Students use information provided as text and empirical data to create a PowerPoint presentation for their class.
- Help Desk (Grade 6 & 10). Students play the role of providing general advice on a community Help Desk and complete some formatting tasks in Word, PowerPoint and Excel.

---

<sup>1</sup> The assessment instrument integrated software from four different providers on a Microsoft Windows XT platform. The two key components of the software package were developed by SkillCheck Inc. (Boston, MA) and SoNet Software (Melbourne, Australia). The SkillCheck system provided the software responsible for delivering the assessment items and capturing student data. The SkillCheck system also provided the simulation, short constructed response and multiple choice item platforms. The SoNet software enabled live software applications (such as Microsoft Word) to be run within the global assessment environment and for the resultant student products to be saved for later grading.

The ICT literacy assessment was administered through a computer environment using sets of six networked laptop computers with all necessary software installed. A total of 3,746 Grade 6 and 3,647 Grade 10 students completed the survey in 263 elementary and 257 secondary schools across Australia. The assessment model defined a single variable, ICT literacy, which integrated three related strands. The calibration provided a high person separation index of 0.93 and a difference in the mean Grade 6 ability compared to the mean Grade 10 ability being of the order of 1.7 logits, meaning that the assessment materials worked well in measuring individual students and in revealing differences associated with a developmental progression.

Describing the scale of achievement involved a detailed expert analysis of the ICT skills and knowledge required to achieve each score level on each item in the empirical scale. Each item, or partial credit item category, was then added to the empirical item scale to generate a detailed, descriptive ICT literacy scale. Descriptions were completed to describe the substantive ICT literacy content within each level.

At the bottom level (1) student performance was described as: Students performed basic tasks using computers and software. They implement the most commonly used file management and software commands when instructed. They recognize the most commonly used ICT terminology and functions.

At the middle level (3): Students working at level 3 generate simple general search questions and select the best information source to meet a specific purpose. They retrieve information from given electronic sources to answer specific, concrete questions. They assemble information in a provided simple linear order to create information products. They use conventionally recognized software commands to edit and reformat information products. They recognize common examples in which ICT misuse may occur and suggest ways of avoiding them.

At the second top level (5): Students working at level 5 evaluate the credibility of information from electronic sources and select the most relevant information to use for a specific communicative purpose. They create information products that show evidence of planning and technical competence. They use software features to reshape and present information graphically consistent with presentation conventions. They design information products that combine different elements and accurately represent their source data. They use available software features to enhance the appearance of their information products.


In addition to providing an assessment of ICT literacy based the national survey gathered information about a range of social characteristics and their access to ICT resources. There was a significant difference according to family socioeconomic status with students whose parents were senior managers and professionals scoring rather higher than those whose parent were unskilled manual and office workers. Aboriginal and Torres Strait Islander students scored lower than other students. There was also a significant difference by geographic location. Allowing for all these differences in background it was found that the computer familiarity had an influence on ICT literacy. There was a net difference associated with frequency of computer use and with length of time for which computers had been used.

The assessment instrument used in 2008 was linked to that used in 2005 by the inclusion of three common modules (including the general skills test) but added four new modules. The new modules included tasks associated with more interactive forms of communication and assessed issues involving responsible use more extensively. In addition the applications functions were based on Open Office.

### *Technology-based assessments in Asia*

In the major economies in Asia, there has been a strong move towards curriculum and pedagogical changes to prepare students for the knowledge economy since the turn of the millennium (Plomp et al., 2009). For example, "Thinking Schools Learning Nation" was the educational focus for

### Planning a trip




You and your classmates are asked to form a group of three to do a project about 'Planning a trip for your grandfather and grandmother'. It is the first time that your grandfather and grandmother visit Hong Kong and they would like to go to some special and distinctive scenic spots in Hong Kong. You and your groupmates are required to suggest 2 scenic spots in Hong Kong and make a presentation for the whole class to present why you have chosen these scenic spots.

**Question 1.1**  
Which search engine(s) (e.g. Yahoo! Hong Kong and Google) have you used for searching on the internet?

**Question 1.2**  
What are the keywords which you have used for you searching?

### Managing information



**Question 2**  
Your groupmate, Susan, also finds some information from the internet and puts them to a Word document. However, she is not familiar with the software and she would like you to make some changes for her document according to the following sample document (required formats have been highlighted), so that the information can be better organized. Apart from the above required changes, You may also use your own ideas to make some changes for her document in order to make it more presentable. (10 mins)

**Victoria Peak**

Victoria Peak is a mountain in the southern of Hong Kong Island. With an altitude of 552 m, it is the highest mountain on the island and the 10th highest in Hong Kong. With about 400,000 visitors every year, the Peak is the biggest tourist attraction in Hong Kong. It offers spectacular views of the city and bay. The Peak area, covering the Peak, Victoria Gap, Mount Soler, Jardine's Corner, Midas, Gough, Robinson Road, is also known as one of Hong Kong's most beautiful landmarks.

**Highlights:-**

- Peak Tram
- Look Po's Star Deck
- Peak Lookout Restaurant

1. Add Susan's name in the header and align it to the right.

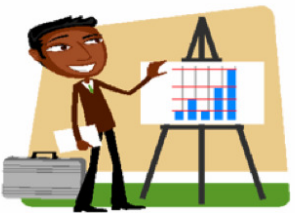
2. Bold and underline the title.

3. Justify the paragraph and change the colour of the selected text blue.

4. Insert a related image here.

5. Add bullet points to the list of areas.

### Presentation



**Question 3**  
You and your groupmates agreed to use the two scenic spots you suggested and decided to create some PowerPoint slides for the presentation with the following structure: (15 mins)

### Discussion

**Question 4**  
You would like to discuss about the scenic spots you suggested with your classmates through the discussion forum. In the forum, you may discuss with your classmates why you suggest these two scenic spots and give opinions to those scenic spots suggested by your classmates. (5 mins)

Click here to start the discussion


Figure 1: Overview of assessment items for technology literacy

Singapore's first IT in Education Masterplan (Singapore MOE, 1997). The Hong Kong SAR government launched a comprehensive curriculum reform in 2000 (EMB, 2001) focusing on developing students' lifelong learning capacity, which is also the focus of Japan's e-learning strategy (Sakayauchi, Maruyama and Watanabe, 2009). Pelgrum (2008) reports a shift in reported pedagogical practice from traditional orientation towards 21st century orientation in these countries between 1998 and 2006, which may reflect the impact of the education policy implementation in these countries.

The focus on innovation in curriculum and pedagogy in these Asian economies may have been accompanied by changes in the focus and format in assessment practice, including high stake examinations. For example, in Hong Kong, a teacher-assessed year-long independent enquiry is being introduced in the compulsory subject Liberal Studies, which forms 20% of the subject score in the school leaving diploma at the end of grade 12, and included in the application for university admission. This new form of assessment is designed to measure the generic skills, which are considered important for the 21st century. On the other hand, technology-based assessment as a means of assessment delivery has not been a focus of development in any of the Asian countries at the system level, even though there may be small-scale explorations by individual researchers. Technology-based assessment innovation is rare. One instance of such is the project on performance assessment of students' information literacy skills conducted in Hong Kong in 2007 as part of the evaluation of the second IT in education strategy in Hong Kong (Law et al., 2007). This Information Literacy Performance Assessment project (ILPA for short, see <http://il.cite.hku.hk/index.php>) is described in some details here as it attempts to use technology in the fourth and fifth domains of assessment as described in an earlier section (whether someone is capable of achieving a higher level of performance with the appropriate use of general or domain-specific technology tools, and the ability to use technology to support collaboration and knowledge building).


Within the framework of the ILPA project, ICT literacy (IL) is not the same as technical competence. In other words, just being technologically confident does not automatically lead to critical and skillful use of information. Technical know-how by itself is inadequate; individuals must possess the

**A day trip to the Hong Kong Ocean Park**

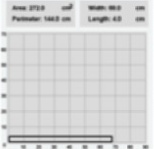


**Question 1**  
John and his family would like to visit the Hong Kong Ocean Park this Saturday. Below is the information about his family. They do not have Smart Fun Annual Pass. Please go to the Internet, find out the ticket prices and help him to calculate the total amount for the whole family's general admission for one day at the park. (10 mins)

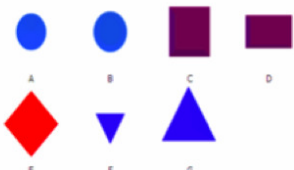
**Question 2**  
Inside the gift shop, Mary wanted to make a pair of identical earrings in the shape of a Christmas tree as a Christmas gift for her mother. She had bought 20cm materials and would use the following software to help her in designing one of the earrings. Can you help her to complete the task? Please write down the length of materials for the pair of earrings. Please show your steps, give the answer and press the "Save" button (8 mins)



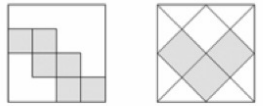
**Question 3**  
Mary bought 144cm ribbon from the gift shop to decorate a rectangular photo frame.  
Use the mouse to drag the rectangle and observe the changes of the length and the width of the rectangle with perimeter 144 cm. (10 mins)



**Question 4**  
At 3:15pm, John's family walked into the Bayview Restaurant for afternoon tea. John's grandfather noticed that there were foods made in different shapes and some of them could be categorized into pyramids or prisms. Can you help John's grandfather to classify the items below into pyramids or prisms? (You may rotate the figures). Please classify the items below into pyramids or prisms (5 mins)



**Question 5**  
They went to the "Sea Jelly Spectacular" and notice that there were two pictures on the wall as shown Figure 1 and 2.  
With reference to the shaded parts, which picture represent(s) 3/8?



**Question 6**  
In the evening, John's father and grandfather were preparing to take the Citybus from the Ocean Park (Aberdeen Tunnel Toll Plaza) to Causeway Bay (Moreton Terrace). They have two options: they could either take route 72A or 76.  
Please find out the details of the Saturday bus fares from the Internet. (6 mins)

**Question 6.1**  
The terminus of two routes are in Causeway Bay (Moreton Terrace). Please find out the starting points of the two routes.

**Figure 2: Overview of grade 5 assessment items for information literacy in mathematics**

cognitive skills needed to identify and address various information needs and problems. ICT literacy includes both cognitive and technical proficiency. Cognitive Proficiency refers to the desired foundational skills of everyday life at school, at home, and at work. Seven Information Literacy Dimensions were included in the assessment:

- Define – Using ICT tools to identify and appropriately represent information needs,
- Access – Collecting and / or retrieving information in digital environments,
- Manage – Using ICT tools to apply an existing organizational or classification scheme for information,
- Integrate – Interpreting and representing information, such as by using ICT tools to synthesize, summarize, compare and contrast information from multiple sources,
- Create – Adapting, applying, designing or inventing information in ICT environments,
- Communicate – Communicating information properly in its context (audience and media) in ICT environments,
- Evaluate - Judging the degree to which information satisfies the needs of the task in ICT environments, including determining authority, bias and timeliness of materials

While these dimensions are generic, a student's IL achievement is expected to be dependent on the subject matter domain context in which the assessment is conducted since the tools and problems may be very different. In this Hong Kong study, the target population included primary 5 (P5, equivalent to grade 5) and secondary 2 (S2, equivalent to grade 8) students in the 2006/07 academic year participated in the assessment. Three performance assessments were designed and administered at each of these two grade levels. At P5, the assessment administered were a generic technical literacy assessment, IL in Chinese language and IL in Mathematics. At S2, they were a generic technical literacy assessment, IL in Chinese language and IL in Science. The generic technical literacy assessment tasks were designed to be the same at P5 and S2 levels as it was expected that personal and family background characteristics may have a stronger influence on a student's technical literacy than age. The assessment tasks for IL in Chinese language were designed to be different as the language literacy for these two levels of students were quite

different. Overviews of the performance assessments are presented for technology literacy in Figure 1, information literacy in Mathematics at grade 5 in Figure 2 and information literacy in Science at grade 8 in Figure 3. It can be seen that the tasks are designed to be authentic, i.e. related to everyday problems that students can understand and care about. Also, subject specific tools are included such as the use of tools to support geometrical manipulation and scientific simulation tools are included for the assessment in Mathematics and Science respectively.

Since the use of technology is crucial to the assessment of information literacy, decisions on what kind of technology and how it is deployed in the performance assessment process is critical. It is important to ensure that students in all schools can have access to a uniform computing environment for the valid comparison of achievement in performance tasks involving the use of ICT. All primary and secondary schools in Hong Kong has at least one computer laboratory where all machines are connected to the Internet. However, the capability, age and conditions of the computers in those laboratories differ enormously across different schools. The assumption of a computer platform that is generic enough to ensure that the educational applications designed can actually be installed in all schools is virtually impossible because of the complexity and diversity of ICT infrastructure in local schools. This problem is further aggravated by the lack of technical expertise in some schools such that there are often a lot of restrictions imposed on the functionalities available to students such as disabling the right-click key which will make some educational applications non-operable, and the absence of common plug-ins and applications such as Active-X and Java runtime engines so that many educational applications cannot be executed. In addition, many technical assistants are not able to troubleshoot to identify problems when difficulties occur.

The need for uniformity is particularly acute in the case of assessing students' task performance using a variety of digital tools. Without a uniform technology platform in terms of the network connections and tools available, it is not possible to conduct fair assessment of students' performance, a task that is becoming increasingly important so as to provide authentic assessment of students' ability to perform tasks in different subject areas that can make use of digital technology. Also, conducting the assessment in the students' own school setting was considered an important criterion as the study also wanted the experience to inform school-based performance assessment.

In order to solve this problem, the Project Team decided on the use of a remote server system - the Microsoft Windows Terminal Server (WTS), after much exploration. This requires the computers in participating schools to be only used as thin clients, i.e. dumb terminals, during the assessment process. It provides a unique and identical Windows' environment for every single user. Every computer in each participating school can log into the system and be used in the same way. In short, all the operations are independent for each client user and functionalities are managed from the server operating system. Students and teachers can take part in learning sessions, surveys or assessments at anytime and anywhere without worrying about the configurations of the computers from which they work. In addition to independent self-learning, collaborative learning with discussion can also be conducted within the WTS.

All student actions made during the assessment process were logged and all their answers stored on the server. Objective answers were automatically scored while the open-ended answers and digital artifacts produced by students were scored online based on a carefully prepared and validated rubric that describes the performance observed at each level of achievement by experienced teachers in the relevant subject domains.

### *Examples of research and development on technology-based assessments in Europe*

Using technology to make assessment more efficient receives a growing attention in several European countries, and a research and development unit of the European Union also facilitates these attempts by coordinating the efforts and organizing workshops (Scheuermann & Björnsson, 2009; Scheuermann & Pereira, 2008).

**Endangered Species**

**Question 2**  
Some new plants and animals have been sent to the Kadoorie Farm for adding to their collection. Your classmates used a digital camera to take pictures of these new species as shown below.  
Among the animals, there is an endangered species found in Hong Kong. An endangered species is a species whose number is so small that it is at the risk of extinction. Can you identify which species it is and what kind of habitat it needs? Use the links on the right hand side or other websites to look for information about endangered species in Hong Kong. (5 mins)

**Visit to the Kadoorie Farm**

**Question 1**  
Before the trip, your teacher has asked you to find a 'Nature Walk Self-guided Map' that you and your classmates can use to take a self-guided tour of Kadoorie Farm during the field trip. (8 mins)

**Question 3.1**  
Please construct a classification diagram:  
1. Classify the above new animals and plants into four suitable categories. Please make reference to the "Nature Walk Self-guided Map" to help the farm staff decide where to place the above newcomers.

**Question 4-7**  
**Part II Exploring Ecosystems Using a Simulation Programme**

Doing ecological experiments may bring danger to the environment. Therefore, scientists often use simulations to find out how changes will affect different species. Now you are going to use one simulation programme to find out more about a pond ecosystem (池塘生態系統). There are totally four questions in the simulation programme. (20 mins)

Please read the instruction carefully once the simulation is started.

Please click the following link to start the simulation programme.

[Start the Simulation Programme](#)

**Figure 3: Overview of grade 8 assessment items for information literacy in science**

At national level, Luxemburg leads the way by introducing a nationwide assessment system, immediately using online testing, while skipping the paper-based step. The current version of the system is able to assess an entire cohort simultaneously. It includes an advanced statistical analysis unit and the automatic generation of the feedback to the teachers. (Plichart, Jadoul, Vandenabeele, & Latour, 2004; Plichart, Latour, Busana, & Martin, 2008). Created, developed, and maintained in Luxemburg by the University of Luxemburg and the Public Research Center Henri Tudor, the core of the TAO (the acronym for Testing Assisté par Ordinateur, the French expression for Computer Based Testing) platform has also been used in several international assessment programs, including the Electronic Reading Assessment (ERA) in PISA 2009 (OECD, 2008), and the OECD Program for International Assessment of Adult Competencies (PIAAC). To fulfill the needs of the PIAAC household survey, Computer-Assisted Personal Interview (CAPI) functionalities have been fully integrated into the assessment capabilities. Several countries have also specialized and further developed extension components that integrate with the TAO platform.

In Germany, a research unit of the *Deutsches Institut für Internationale Pädagogische Forschung* (DIPF, German Institute for International Educational Research, Frankfurt) launched a major project that adapts and further develops the TAO platform. "The main objective of the 'Technology Based Assessment' (TBA) project at the DIPF is to establish a national standard for technology-assisted testing on the basis of innovative research and development according to international standards as well as reliable service."<sup>2</sup> The technological aspects of the developmental work include an item-builder software, the creation of innovative item formats (e.g. complex and interactive contents), feedback routines, and computerized adaptive testing and item banks. Another innovative application of TBA is the measurement of complex problem solving abilities. The related experiments began in the late 1990s, and a large-scale assessment was conducted in the framework of the German extension of PISA 2003. The core of the assessment software is a finite automaton, which can be easily scaled in terms of item difficulty, and can be realized in a number of contexts (covers stories, "skins"). This approach provided an instrument that measures a cognitive construct separable both from analytical problem solving and general intelligence (Wirth & Klieme, 2003, Wirth & Funke, 2005). The most recent and a more sophisticated tool uses the MicroDYN

<sup>2</sup> See [http://www.tba.dipf.de/index.php?option=com\\_content&task=view&id=25&Itemid=33](http://www.tba.dipf.de/index.php?option=com_content&task=view&id=25&Itemid=33) for the mission statement of the research unit:

approach, where the testee faces a dynamically changing environment. (Blech & Funke, 2005; Greiff, & Funke, 2008). One of the major educational research initiative, the Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes<sup>3</sup> also includes several TBA related studies (e.g. dynamic problem solving, dynamic testing, rule-based item generation).

In Hungary, the first major technology-based testing took place in 2008. An inductive reasoning test was administered to a larger sample of 7<sup>th</sup> grader students both in paper and pencil version and online (using the TAO platform) to examine the media effects. The first results indicate, that, although the global achievements highly correlate, there are items with significantly different difficulties in the two media, and there are persons, who are significantly better on one of the media (Csapó, Molnár, & R. Tóth, 2009). In 2009, a large-scale project was launched to develop an online diagnostic assessment system for the first six grades of primary school in reading, mathematics and science. The project includes the development of assessment frameworks, devising a large number of items both on paper and on computer, building item banks, using technologies for migrating items from paper to computer, and research on comparing the achievements on the tests using different media.

### *Examples of technology in assessment in the US*

In the US, there are many instances in which technology is being used in large-scale summative testing. At the primary and secondary levels, the largest technology-based testing programs are the Measures of Academic Progress (Northwest Evaluation Association), the Virginia Standards of Learning Tests (Virginia Department of Education), and the Oregon Assessment of Knowledge and Skills (Oregon Department of Education). The Measures of Academic Progress (MAP) is a computer-adaptive test series offered in reading, mathematics, language usage, and science at the primary and secondary levels. MAP is used by thousands of school districts. The test is linked to a diagnostic framework, DesCartes, which anchors the MAP score scale in skill descriptions that are popular with teachers because they appear to offer formative information. The Virginia Standards of Learning (SOL) Tests are a series of assessments that cover reading, mathematics, sciences, and other subjects at the primary and secondary levels. Over 1.5 million SOL tests are taken online annually. The Oregon Assessment of Knowledge and Skills (OAKS) is an adaptive test in reading, mathematics, and science in primary and secondary grades. The OAKS is approved for use under *No Child Left Behind*, the only adaptive test with that status. OAKS and those Virginia SOL tests used for *NCLB* purposes have high stakes for schools because sanctions can be levied for persistently poor test performance. Some of the tests may also have considerable stakes for students, including those measures that factor into end-of-course grading, promotion, or graduation decisions. MAP, OAKS, and SOL online assessments are believed to be exclusively multiple-choice tests.

Online tests offered by the major test publishers for what the publishers describe as formative assessment purposes include Acuity (CTB/McGraw-Hill) and the PASeries (Pearson). Perhaps more aligned with current concepts of formative assessment are the Cognitive Tutors (Carnegie Learning). The Cognitive Tutors, which focus on algebra and geometry, present problems to students, use their responses to dynamically judge understanding, and then adjust instruction accordingly.

At the postsecondary level, ACCUPLACER (College Board) and COMPASS (ACT) are summative tests used for placing entering freshmen in developmental reading, writing, and mathematics courses. All sections of the tests are adaptive, except for the essay which is automatically scored. The tests have relatively low stakes for students. The Graduate Record Examinations (GRE) General Test (ETS), the Graduate Management Admission Test (GMAT) (GMAC), and the Test of English as a Foreign Language (TOEFL) iBT (ETS) are all offered on computer. All three summative tests are high-stakes ones used in educational admissions. Sections of the GRE and

<sup>3</sup> See <http://kompetenzmodelle.dipf.de/en/projects>



GMAT are multiple-choice, adaptive tests. The writing sections of all three tests include essays, which are scored automatically, as well as by one or more human graders. The TOEFL iBT also has a constructed-response speaking section, with digitized recordings of examinee responses scored by human judges. A formative assessment, TOEFL Practice Online (ETS), includes speaking questions that are scored automatically.

### *Applying technology in international assessment programs*

The large-scale international assessment programs currently in operation have their origins in the formation of the *International Association for the Evaluation of Educational Achievement (IEA)* in 1958. The formation of the IEA arose from a desire to focus comparative education on the study of variations in educational outcomes such as knowledge, understanding, attitude and participation as well as the inputs to education and the organization of schooling. Most of the current large-scale international assessment programs are conducted by the IEA and the *Organization for Economic Co-operation and Development (OECD)*.

The IEA has conducted the *Trends in International Mathematics and Science Study (TIMSS)* at Grade 4 and Grade 8 every four years since 1995 and has its fifth cycle scheduled for 2011 (Mullis, Martin, & Foy, 2008; Martin, Mullis, & Foy, 2008). It has also conducted the *Progress in International Reading Literacy Study (PIRLS)* at Grade 4 every five years since 2001 and has its third cycle scheduled for 2011 (Mullis, Martin, Kennedy, & Foy, 2007). In addition the IEA has conducted periodic assessments in *Civic and Citizenship Education (ICCS)* in 1999 (Torney-Purta, Lehmann, Oswald, & Schulz, 2001) and 2009 (Schulz, Fraillon, Ainley, Losito, & Kerr, 2008) and is planning an assessment of *Computer and Information Literacy (ICILS)* for 2013 .

The OECD has conducted the *Programme for International Student Assessment (PISA)* among 15-year-old students every three years since 2000 and has its fifth cycle scheduled for 2012 (OECD, 2007). It assesses Reading, Mathematical and Scientific Literacy in each cycle but with one of those three being the major domain in each cycle. In the 2003 cycle it included an assessment of Problem Solving. The OECD is also planning to conduct a *Programme for the International Assessment for Adult Competencies (PIAAC)* in 2011 in 27 countries. The target population is adults aged between 16 and 65 years and each national sample will be a minimum of 5,000 people, who will be surveyed in their homes (OECD, 2008). It is designed to assess literacy, numeracy and “problem solving skills in technology-rich environments” as well as surveying how those skills are used at home, work and in the community.

TIMSS and PIRLS have made use of ICT for web based school and teacher surveys but have not yet made extensive use of ICT for student assessment. An international option of web-based reading was planned to be part of PIRLS 2011 and modules were developed and piloted. Whether the option proceeds to the main survey will depend upon the number of countries opting to include the module. The planned *International Computer and Information Literacy Study (ICILS)* will examine the outcomes of student computer and information literacy (CIL) education across countries. It will investigate the variation in CIL outcomes between countries, and between schools within countries, so that those variations can be related to the way CIL education is provided. CIL is envisaged as the capacity to use computers to investigate, create and communicate in order to participate effectively at home, at school, in the workplace and in the community. It brings together computer competence and information literacy and envisages the strands of accessing and evaluating information, as well as producing and exchanging information. In addition to a computer-based student assessment the study will include computer-based student, teacher and school surveys. It will also incorporate a national contexts survey.

PISA has begun to use ICT in the assessment of the domains it assesses. In 2006 PISA scientific literacy was the major domain and the assessment included an international option entitled a *Computer-Based Assessment of Science (CBAS)*. CBAS was delivered by a Test Administrator taking a set of six laptop computers to each school with the assessment system installed on a wireless or cabled network, with one of the networked PCs acting as an administrator's console.

Student responses were saved during the test both on the student's computer and on the Test Administrator's computer. An online translation management system was developed to manage the translation and verification process for CBAS items. A typical CBAS item consisted of a stimulus area, containing text and a movie or flash animation, and a task area containing a simple or complex multiple-choice question, with radio buttons for selecting the answer(s). Some stimuli were interactive, with students able to set parameters by keying-in values or dragging scale pointers. There were a few drag-and-drop tasks, and some multiple-choice questions required students to select from a set of movies or animations. There were no constructed response items, all items were computer-scored, and all student interactions with items were logged. CBAS field trials were conducted in 13 countries but the option was included in the main study in only three countries.

PISA 2009 has reading literacy as a major domain and included an *Electronic Reading Assessment* (ERA) as an international option. The ERA test uses a test administration system (TAO) developed through the University of Luxembourg. TAO can deliver tests over the internet, across a network or (as is the case with ERA) on a standalone computer with student responses collected on a memory (USB) stick. The ERA system includes an online translation management system, and an online coding system for free-response items. An ERA item consists of a stimulus area that is a simulated multi-page web environment, and a task area. A typical ERA item involves students navigating around the web environment to answer a multiple-choice or free-response question. Other types of tasks require students to interact in the stimulus area by clicking on a specific link, making a selection from a drop-down menu, posting a blog entry or typing an email. Answers to constructed-response items are collated for marking by humans and other tasks are scored by computer. The PISA 2009 *Reading Framework* is currently being produced and it will articulate the constructs assessed in the ERA and relationship of those constructs to the paper-based assessment. Subsequent cycles of PISA plan to make further use of computer-based assessment.

PIAAC builds on previous international surveys of adult literacy (such as IALS and ALL) but is extending the range of competencies assessed and investigating the way skills are used at work. Its assessment focus is on literacy, numeracy, reading components and "problem solving in technology-rich environments" (OECD, 2008). "Problem solving in technology-rich environments" refers to the cognitive skills required in the information age rather than computer skills and similar to what is often called information literacy. This aspect of the assessment will focus on solving problems using multiple sources of information on a laptop computer. The problems are intended to involve accessing, evaluating, retrieving and processing information and incorporate technological and cognitive demands. The conceptions of literacy and numeracy in PIAAC emphasize competencies situated in a range of contexts and application, interpretation and communication. The term "reading components" refers to basic skills such as "word recognition, decoding skills, vocabulary knowledge and fluency" (OECD, 2008). In addition to assessing these domains PIAAC surveys adults in employment about the types and levels of a number the general skills used in their workplaces as well as background information. This background information includes data about how they use literacy, numeracy, and technology skills in their daily lives, their education background, employment experience and demographic characteristics (OECD, 2008). The assessment, and the survey, is computer-based and administered to people in their homes by trained interviewers. The assessment is based on the TAO system.

In international assessment programs, as in national and local programs, two themes in the application of ICT are evident. One is the use of ICT to assess better the domains that have traditionally been the focus of assessment in schools: reading mathematics and science. "Assessing better" means using richer and more interactive assessment materials, using those materials to assess aspects of the domains that have been hard to assess and possibly extending the boundaries of those domains. This theme is evident in the application of ICT thus far in PISA and PIRLS. A second theme is the use of ICT to assess more generic competencies. This is evident in the proposed ICILS and the imminent PIAAC which both propose to assess the use of computer technology to assess a broad set of generalisable and transferable knowledge, skills and understandings that are used to manage and communicate information. They are dealing with the intersection of technology and information literacy (Catts and Lau 2008).

## **Technology for item development and test management**

One of the main success factors in developing a modern technology-based assessment platform is certainly not located at the technological level, but instead the iterative and participatory design mode that should be adopted in the platform design and development process. Indeed, as often observed in the field of scientific computing, the classical customer-supplier relationship from a pure Software Engineering service point of view is highly ineffective in such a dramatically complex problem where Computer Science considerations are sometimes not distinguishable from psychometric considerations. On the contrary, a successful technology-based assessment (TBA) expertise must be built by deep immersion of both disciplines.

In addition to the trans-disciplinary approach, two other factors are also increasing the chances to fulfill the needs for the assessment of the 21st century skills. First, the platform should be designed and implemented independently from specific context of use. This requires a more abstract level of design leading to high-level and generic requirements that might look distant from concrete user or pragmatic organizational process concepts. Therefore, a strong commitment and understanding on this issue from the assessment experts together with a thorough understanding of the TBA domain and good communication from the technologists are essential. As already stressed in e-learning contexts, a strong collaboration between disciplines is essential (Corbiere 2008).

Secondly, TBA processes and requirements are highly multiform and convey a tremendous diversity of needs and practices in the education domain (Martin, Busana & Latour 2009) but also more generally when ranging across assessment classification descriptors, *i.e.*, from researchers in psychometrics, educational measurement, or experimental psychology to large-scale assessment and monitoring professionals, or from education context to human resource management. As a consequence, willing to build a comprehensive and detailed a priori description of the needs appears totally illusive. On the contrary, both assessment and technology experts should acknowledge the need to iteratively elicit the context specific requirements that are further abstracted in the analysis phase while developing the software in a parallel process and in such a way that unexpected new features can be added with the least impact on the code. This process is likely to be the most efficient way to tackle the challenge.

### **Principles for developing technological platforms**

#### *Enabling assessment of reliability of data and versatility of instruments*

While strongly depending on providers' business models, the open-source paradigm in this area bears two fundamental advantages. The full availability of the source does not only enable the possibility to assess the implementation and reliability of the measurement instruments (which is a crucial aspect of scientific computing in general and psychometrics in particular), but also to fine-tune the software to very specific needs and contexts with a full control of the implementation process and costs, while benefiting from the contributions of a possibly large community of users and developers. Built-in extension mechanisms enable developers from within the community to create new extensions and adaptations without modifying the core layers of the application, and to share their contributions.

#### *Enabling efficient management of assessment resources*

An integrated technology-based assessment should enable an efficient management of assessment resources (items, tests, subjects and groups of subjects, results, surveys, deliveries...) and provide support to the organizational processes (depending on the context, translation and verification for instance); the platform should also ensure delivering the cognitive instruments and background questionnaires to the test takers and possibly other stakeholder, together with collecting, post-processing and exporting results and behavioral data. In order to support complex collaborative

processes such as those of large-scale international surveys, modern CBA platform should offer annotation with semantically rich meta-data as well as collaborative capabilities.

Complementary to the delivery of Cognitive Instruments, modern CBA platforms should also provide a full set of functionalities to collect background information, mostly about the test taker, but also possibly about any kind of resources involved in the process. As an example, in the PIAAC survey, the Background Questionnaire (consisting of questions, variables and logical flow of questions with branching rules) has then been fully integrated into the global survey workflow, together with the cognitive instrument booklet.

In the ideal case, interview items, assessment items, and entire tests or booklets are interchangeable. As a consequence, very complex assessment instruments can be designed fully integrating cognitive assessment and Background Data Collection in a single flow, on a unique platform.

### *Accommodating a diversity of assessment situations*

In order to accommodate the large diversity of assessment situations, modern computer-assessment platforms should enable a large set of deployment modes from full web-based deployment on a large server-farm with load balancing enabling the delivery of a large number of simultaneous tests to CD's or memory sticks ran on school desktops. As an illustration, the latter solution has been used in the PISA ERA 2009. In the PIAAC international survey, the deployment has been made using a Virtual Machine installed on individual laptops carried out by interviewers into the participating households. In classroom contexts, wireless Local Area Network (LAN) using a simple laptop as server and tablet PC's as client machines used by the test takers can also be used.

## **Item building tools**

### *Balancing usability and flexibility*

Item authoring is one of the crucial tasks in the delivery of technology-based assessments. So far, depending on the requirements induced by the frameworks, different strategies have been pursued, ranging from hard-coded development by software programmers to easy-to-use simple template-based authoring. Even if it seems intuitively the most natural solution, the purely programmatic process should in general be avoided. Such outsourcing strategy (disconnecting the content specialists from the software developers) usually requires very precise specifications that most often item designer and framework expert are not familiar with. In addition, it lengthens the timeline, reduces the number of iterations preventing trial and error procedures. Moreover, this process does not scale well when the number of versions of every single item increases, as is the case when one has to deal with many languages and country-specific adaptations. Of course, there will always be a tradeoff between usability and easiness (that introduces strong constraints and low freedom in the item functionalities) and flexibility to describe rich interactive behaviors (that introduces a higher level of complexity when using the tool). In most situations, it is advisable to enable different interfaces dedicated to users with different levels of IT competency. To face the challenge of allowing great flexibility while keeping the system usable with a minimum of learning, template-driven authoring tools built on a generic expressive system are probably one of the most promising technologies. Indeed, it enables, using a single system to hide inherent complexity when building simple items while letting more power users the possibility to further edit advances features.

### *Separating item design and item implementation*

Item-authoring processes can be further subdivided in different tasks such as item design (setting up the item content, task definition, response domain, and possibly scenarios) and item implementation (translating the item design into the computer platform so that the item becomes an

executable piece of software). Depending on the complexity of the framework, different tools can be used to perform both tasks. In some circumstances, building the items iteratively enables one to keep managing the items complexity. This can be achieved by first creating a document describing all the details of the item scenario, based on the framework definition, and then transforming the document into an initial implementation template or draft. An IT specialist or a trained power user then further expands the implementation draft to produce the executable form of the item. This process more effectively addresses the stakeholders by remaining as close as possible to the user usual practice. Indeed, modern web- and XML-based technologies, such as CSS (Lie & Bos, 2008), Javascript, HTML (Raggett, Le Hors & Jacobs, 1999), XSLT (Kay, 2007), Xpath (Berglund *et al.*, 2007), and Xtiger (Kia, Quint & Vatton, 2008) among others allow easily building template-driven authoring tools (Flores, Quint & Vatton, 2006), letting the user having the same experience as when editing a word document. The main difference with respect to editing a word document is that the information is structured with respect to concepts pertaining to the assessment and framework domains. This feature enables automatic transformation of the item design into a first draft-implemented version that can be passed to another actor in the item production process.

### *Distinguishing authoring from run-time and management platform technologies*

It has become common practice in the e-learning community to strictly separate the platform component, from the learning content and the tools used to design and execute the learning content. TBA is now starting to follow the same trend. However, practices inherited from paper-and-pencil assessment as well as additional complexity arising from psychometric constraints and models, sophisticated scoring and new advanced framework has somehow slowed down the adoption of this concept. In addition, the level of integration of IT and psychometricians remains low in the community. This often leads to a lack of more global or systemic vision on both sides. As a result, a significant number of technology-based assessments are implemented following a silo approach centered on the competency to be measured and including in single closed software all the functionalities. Whenever the construct, the framework or the type of items increases, this model is no longer viable in the long run. On the contrary, the platform approach and the strict separation of test management and delivery layers, together with the strict separation of the item runtimes and authoring, is the only scalable solution with respect to high diversity.

### *Items as interactive composite hypermedia*

In order to fully exploit the most recent advances in computer media technologies, one should be able to combine in an integrative manner various types of interactive media enabling various types of user interactions and functionalities. In the case ubiquity is a strong requirement, i.e., making assessment available everywhere; modern web technologies must be seriously considered. Indeed, even if they still suffer from performance and lack of advanced features that can be found in platform-dedicated tools, they however provide the sufficiently rich set of interaction features that one needs in most assessments. In addition, these technologies are readily available on a wide range of cost-effective hardware platform, together with cost effective licenses, if not open-source. Moreover, web technologies in general enable very diversified types of deployment across networks (their initial vocation), as well as locally on laptops or other devices. This important characteristic makes deployments very cost-effective and customizable with respect to assessment contexts.

This dramatically changes the vision one may have about item authoring tools. Indeed, on one hand, IT developers build many current complex and interactive items programmatically, while on the other hand very simple items with basic interactions and data collection mode such as multiple-choice items are most often built using templates or using simple descriptive languages accessible to non-programmers (like basic HTML).

There are currently no easy and user-friendly intermediate solutions between these two extremes. Yet, most often, and especially when items are built on according to dynamic stepwise scenarios, the item needs to define and control a series of behaviors and user interactions into the item. If we abstract ourselves from the media *per se* (the image, the video, a piece of an animation, or a sound

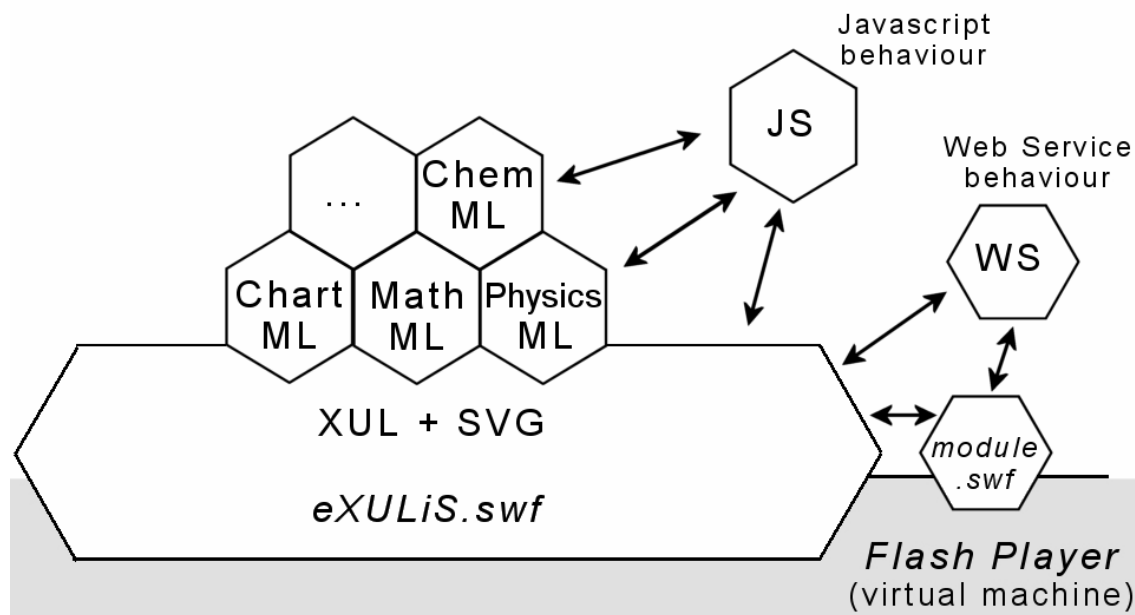
file for instance), one realizes that a large deal of user interactions and system responses can be modeled as change of state driven by events and messages triggered by the user and transmitted between the item objects.

The role of the item developer is to instantiate the framework into a scenario and to translate this scenario into a series of content and testee actions. In paper-and-pencil assessments, expected testee actions are reified in the form of instructions, and the data collection consists uniquely in collecting an input from the test-taker. Since the paper instrument cannot change its state during the assessment, no behavior or response to the user can be embedded in the instrument.

One of the fundamental improvements brought by technology-based assessment is the capacity to embed system responses and behaviors into the instruments, and enabling it to change its state according to the test taker manipulations. Coming back to the instantiation of the framework in a technology-based assessment setting, the reification of expected testee action is no longer in the form of instructions only, but also programmed into interaction patterns between the subject and the instrument, designed in such a way that they should drive the subject towards the expected sequence of actions. In the meantime, one can also collect the history of the user interaction as being part of the input in addition to the explicit information input by the test taker. As a consequence, depending on the framework, richness of the item arises from both the type of media content and the user interaction patterns that drive the state of the whole items and all its components over time.

This clearly separates different concerns from an authoring tool perspective. First, just as if they were manipulating tools to create paper-and-pencil items, item developers must create non- or loosely-interactive media contents in the form of texts, images, or sounds separately. Each of these media encapsulates their own set of functionalities and attributes. Second, they will define the structure of their items both in terms of item logics (stimulus, tasks or questions, response collection...). Third, they will populate the items with the various media they need. And fourth, they will set up the interaction scheme between the user and the media, and between the different media.

Such a high level Model-View-Controller architecture of item authoring tools, based on XML (Bray *et al.*, 2006, 2008) and web-technologies, results in highly cost-effective authoring processes. They are claimed as fostering *wider access to high quality visual interfaces and shorter authoring cycles for multidisciplinary teams* (Chatty *et al.*, 2004). It first let item developer use their favorite authoring tools to design media contents of various types instead of learning complex new environments and paradigms. In most cases, several of these tools are available as open-source software. In addition, the formats manipulated by these are often open-standards available for free from the web community. Then, when considering the constant evolution of assessment domains, constructs, frameworks, and finally instrument specification, one should be able to extend rapidly and easily the scope of interactions and/or type of media that should be encapsulated into the item. Having separated the content from the layout and the behavioral parts, the inclusion of new sophisticated media into the item and their inclusion into the user-system interaction patterns is made very easy and cost-effective. In the field of science, sophisticated media such as molecular structure manipulation and viewer like Jmol (Herráez, 2007), (Willighagen & Howard, 2007) and RasMol (Sayle & Milner-White, 1995; Bernstein, 2000), interactive mathematical tool dedicated to space geometry, or other simulations can be connected to the other parts of the item. Mathematic notations, or 3D scenes described in X3D (Web3D Consortium, 2007, 2008) or MathML (Carlisle, Ion, Mine & Poppelier, 2003) format, respectively and authored with open-source tools, can also be embedded and connected into the interaction patterns of the items, together with SVG (Ferraiolo, Jun & Jackson, 2009) images, and XUL (Mozilla Foundation) or XAML (Microsoft) interface widget for instance. These principles have been implemented in eXULiS (Jadoul, Plichart, Swietlik & Latour, 2006), as illustrated in Figure 4. A conceptually similar but technically different approach where a conceptual model of an interactive media undergoes a series of transformations to produce the final executable has been recently experimented by Tissoires and Conversy (2008).



**Figure 4: Illustration of eXULIS handling & integrating different media types & services**

Going further in the transformational document approach, the document-oriented GUI enables users to edit documents directly on the web, considering that the Graphical User Interface is also a document (Draheim, Lutteroth & Weber, 2006). Coupled with XML technologies and composite hypermedia item structure, this technique enables addressing item authoring as a layered edition of embedded documents describing different components or aspects of the item.

Just as it has been claimed for the assessment resource management level, item authoring will also largely benefit from being viewed as interactive hypermedia integration platforms. As for the management platform, such a horizontal approach guarantees cost-effectively, time-effectively, openness, and flexibility, while keeping the authoring complexity reasonable.

#### *Extending item functionalities with external on-demand services*

The definition of item behavior and user interaction patterns presented here above covers a large part of the item functional space. Composite interactive hypermedia can indeed fulfill most of the simple interactions that control the change of state of the item according to the user actions. However, there exist domains where more complex computations are expected at test time, *i.e.*, during the test administration. Schematically, one can distinguish four classes of such situations: when an automatic feedback to the test taker is needed (mostly in formative assessments); when an automatic scoring is expected for complex items; when using advanced theoretical foundation such as Item Response Theory and adaptive testing; and finally when the domain requires complex and very specific computation to drive the item change of state (simulation, science).

When items are considered in a programmatic way, *i.e.*, as a closed piece of software created by programmers, or when items are created from specialized software templates, these issues are dealt with at design and software implementation time. As a consequence, the complex computations are built-in functions of the items. Very differently, when considering item as a composition of interactive hypermedia as described here above, this built-in programmatic approach is no longer viable in the long run. The reasons of this lack of viability are twofold. First, from a computational cost point of view, the execution of these complex dedicated functions may be time-consuming. If items are based on web technologies and are client-oriented (the execution of the item functionalities is done on the client – the browser – rather than on the server), this may lead to

problematic lag-time between the action of the user and the computer response. More than being an ergonomic and user comfort issue, this may seriously endanger the quality of collected data. Second, from a cost and timeline point of view, proceeding in such a way implies lower reusability of components across domains, and then higher development cost, less flexibility, more iteration between the item developer and the programmer, and finally longer delays.

Factorizing these function from the item framework constitutes an obvious solution. From a programmatic approach this would lead to the construction of libraries programmers can reuse when programming the items. In a more interesting, versatile and ubiquitous way, considering these functions as components that fits into the integrative composition of interactive hypermedia bears serious advantages. On one hand, it enables abstracting the functions in the form of high-level software services that can be invoked by the item author (acting as an integrator of hypermedia and designer of user-system interaction patterns); and on the other hand it enables a higher reusability of components across domains. Moreover, in some circumstances mostly depending on the deployment architecture, invocation of externalized software services may also solve partially the computational cost problem.

Once again, when looking at currently available and rapidly evolving technologies, Web technologies and Service-Oriented approaches, based on UDDI (Clement, Hately, von Riegen & Rogers, 2004), WSDL (Booth & Liu, 2007), and SOAP (Gudgin *et al.*, 2007) standards, offer an excellent ground to implement this vision without drastic constraints on the deployment modalities.

The added value of such an approach regarding the externalization of software services can be illustrated in various ways. When looking at new upcoming frameworks and the general trend in education from contents towards more participative inquiry-based learning, together with the globalization and the increase of complexity of our modern societies, one expects that items will also follow the same transformations. Seeking to assess citizen capacity to evolve in a more global and systemic multi-layered environment (as opposed to past local and strongly stratified environments where people only envision the nearby  $n \pm 1$  levels) it seems obvious that constructs, derived frameworks and instantiated instruments and items will progressively bear the characteristics of globalized systems. This constitutes an important challenge for technology-based assessment that must support not only items and scenarios that are deterministic but also new ones that are not deterministic or complex in nature. The complexity in this view is characterized either by a large response space when there exist many possible sub-optimal answers, or by uncountable answers. This situation can typically occur in complex problem solving where the task may refer to multiple concurrent objectives and yield to final solutions that may no be unique and/or consisting as an optimum set of different sub-optimal solutions. Automatic scoring and more importantly the management of system responses requires sophisticated algorithms that must be executed at test-taking time. Embedding such algorithms into the item programming would increase dramatically the development time and cost of items, while lowering the reusability. Another source of complexity in this context that advocates the service approach arise when the interactive stimulus is a non-deterministic simulation (at the system level, not at local level of course). Multi-agent systems (often embedded in modern games) are such typical systems that are best externalized instead of being embarked into the item.

In more classical instances, externalizing IRT algorithms in services invoked from the item at test-taking time will bring a high degree of flexibility to item designers and researchers. Indeed, various item models, global scoring algorithms, and item selection strategies in adaptive testing can be experimented at low cost without modifying the core of existing items and tests. In addition, it enables using existing efficient packages instead of redeveloping the services. Another typical example can be found in science when one may need specific computation of energies or other quantities, or particular simulation of phenomenon. Once again, the service approach takes advantage of the existing efficient software that is available on the market. Last but not least, when assessing software, database, or XML programming skills, some item designs include compilation or code execution feedbacks to the user at test-taking time. One would certainly never incorporate or develop a compiler or code validation into the item. On the contrary, the obvious solution is to call



these tools as services (or web services). This technique has been experimented in assessment following programming trainings for unemployed persons (Jadoul & Mizohata, 2006).

Finally, and to conclude this point, it seems that the integrative approach in item authoring is amongst the most scalable one in terms of time, cost, and item developer accessibility. Following this view, an item becomes a consistent composition of various interactive hypermedia and software services (that might be in turn interactive or not) that have been developed specifically for dedicated purposes and domains but reusable across different situations rather than a closed piece of software or media produced from scratch for a single purpose. This reinforces the so-called horizontal platform approach to the detriment of the current vertical full programmatic silo approach.

### **Item banks, storing item meta-data**

Item banking is often considered as the central element in the set of tools supporting computer-based assessment. They consist in collections of items characterized by meta-data and most often collectively built by a community of item developers. Items in item banks are classified according different aspects such as difficulty, type of skill, or topic (Conole & Waburton, 2005).

A survey on item banks performed in 2004 reveals that most reviewed item banks has been implemented using SQL databases and XML technologies in various way, concerning meta-data, few implemented meta-data beyond the immediate details of items (Cross, 2004a). The two salient metadata frameworks that arose from this study are derived from IEEE LOM (IEEE LTSC, 2002) and IMS QTI (IMS 2006). Since it is not our purpose here to discuss in detail the metadata framework, but rather to discuss some important technologies that might support the management and use of semantically rich metadata and item storage, the interested reader can refer to the IBIS report (Cross, 2004b) for a more detailed discussion about metadata in item banks.

When considering item storage, one should clearly separate the storage of the item *per se*, or its constituting parts, from the storage of metadata. As already quoted by the IBIS report, relational databases remain today the favorite technology. However, with the dramatic uptake of XML-based technologies and considering the current convergence between the document approach and the interactive web application approach around XML formats, dedicated XML database can also be considered.

Computer-based assessment meta-data are used to characterize the different resources occurring in the various management processes, such as subjects and target groups, items and tests, deliveries and possibly results. In addition, in the item authoring process, metadata can also be of great use to facilitate the search and exchange of media resources that will be incorporated into the items. This is of course of high importance when considering the integrative hypermedia approach.

As a general statement, metadata can be used to facilitate

- the item retrieval when creating a test, concentrating on various aspects such as item content, purposes, models, or other assessment qualities (the measurement perspective); the media content perspective (material embedded into the items); the construct perspective; and finally the technical perspective (mostly for interoperability reasons);
- the correct use of item in consistent contexts from the construct perspective and the target population perspective;
- the track of usage history by taking into accounts the contexts of use, in relation to the results (scores, traces and logs);
- the extension of the result exploitation by strengthening and enriching the link with diversified background information stored in the platform; and
- sharing of content and subsequent economies of scales when inter-institutional collaborations are set up.

Different approaches can be envisioned concerning the management of metadata. Very often, metadata are specified in the form of XML manifests that describe the items or other assessment resources. When exchanging, exporting or importing the resource, the manifest is serialized and transported together with the resource (sometimes the manifest is embedded into the resource). Depending on the technologies used to implement the item bank, these manifests are either stored as is or parsed into the database. The later situation implies that the structure of the metadata manifest is reflected into the database structure. This makes the implementation of the item bank dependent on the choice of a given metadata framework, and moreover, that there is a common agreement in the community about the metadata framework, which then constitutes an accepted standard. While highly powerful, valuable and generalized, with regard to the tremendous variability of assessment contexts and needs, one may rapidly experience the “standard curse”, *i.e.*, the fact that there always exists a situation where the standard does not fit to the particular need. In addition, even if this problem can be circumvented, interoperability issue may arise when one wishes to exchange resources with another system built according to another standard.

Starting from our fundamental stance regarding the need for versatile and open platform as the only economically viable way to embrace the assessment diversity and future evolution, a more flexible way to store and manage metadata should be proposed in further platform implementation. Increasing the flexibility in metadata management has two implications: first, the framework (or metadata model, or meta-model) should be made updatable, and second, the data structure should be independent from the metadata model. From an implementation point of view, *i.e.*, the way the metadata storage is organized and the way metadata exploitation functions are implemented, this requires a soft-coding approach instead of traditional hard-coding approach. In order to do so, in a Web-based environment, Semantic Web (Berners-Lee, Hendler, & Lassila, 2001) and ontology technologies are among the most promising technologies. As an example, such approach is under investigation for e-learning platform to enable individual learners to use their own concepts instead of being forced to conform to a potentially inadequate standard (Tan, Yang, Tang, Lin & Zhang, 2008). This enables one to create annotation of Learning Objects using ontologies (Gašević, Jovanović, & Devedžić, 2004). In a more general stance, impacts and issues related to Semantic Web and ontologies in e-learning platforms have been studied by Vargas-Vera and Lytras (Vargas-Vera & Lytras, 2008).

In the Semantic Web vision, web resources are associated with the formal description of their semantics. The purpose of the semantic layer is to enable machine reasoning on the content of the web, in addition to the human processing of documents. Web resource semantics is expressed as annotations of documents and services in metadata that are in turn resources of the Web. The formalism used to annotate Web resources is triple model called Resource Description Framework (RDF) (Klyne & Carroll, 2004), serialized among other syntaxes in XML. The annotations makes reference to a conceptual model called ontology and modeled using RDF Schema (RDFS) (Brickley & Guha, 2004) or Ontology Web language (OWL) (Patel-Schneider, Hayes, & Horrocks, 2004).

The philosophical notion of ontology has been extended in IT to denote the artifact produced after having studied the categories of things that exist or may exist in some domain. As such, an ontology results in a shared conceptualization of things that exist and make up the world or a subset of it, *i.e.*, the domain of interest (Sowa, 2000; Grubber, 1993; Mahalingam & Huns, 1997). An inherent characteristic of ontologies that makes them different from taxonomies is that they bear intrinsically the semantics of the concepts they describe (Grubber, 1991; van der Vet & Mars, 1998; Hendler, 2001; Ram & Park, 2004) with any number of abstraction levels as required. Taxonomies present an external point of view of things, *i.e.*, a convenient way to classify things according to a particular purpose. In a very different fashion, ontologies present an internal point of view of things, *i.e.*, it tries to figure out how things are, as they are, using a representational vocabulary with formal definitions of the meaning of the terms together with a set of formal axioms that constrain the interpretation of these terms (Maedche & Staab, 2001).

Fundamentally, in the IT field, ontology describes explicitly the structural part of a domain of knowledge in a knowledge-based system. In this context, ‘explicit’ means that there exists some

language with precise primitives (Maedche & Staab, 2001) and associated semantics that is used as a framework for expressing the model (Decker *et al.*, 2000). This ensures that an ontology is machine processable and exchangeable between software or human agents (Guarino & Giaretta, 1995; Cost *et al.*, 2002). In some pragmatic situations, it simply consists in a formal expression of meta-data describing information units (Khang & McLeod, 1998).

Ontology-based annotation framework supported by RDF Knowledge-Based systems enables the management of many evolving metadata framework which conceptual structures are represented in the form of ontologies, together with the instances of these ontologies representing the annotations. In addition, depending on the context, users can also define their own models in order to capture other features of assessment resources that are not considered in the metadata framework. In the Social sciences, such framework is currently used to collaboratively build and discuss models on top of which surveys and assessments are built (Jadoul & Mizohata, 2007).

## **Delivering technologies**

There is a range of methods for delivering computer-based assessments to students in schools and other educational institutions. The choice of delivery method needs to take account of the requirements for the assessment software, the computer resources in schools (numbers, co-location and capacity) and the bandwidth available for school connections to the internet. Key requirements for delivery technologies are that they provide the basis for the assessment to be presented with integrity (uniformly and without delays in imaging), are efficient in the demands placed on resources and are effective in capturing student response data for subsequent analysis<sup>4</sup>.

### **Factors shaping choice of delivery technology**

The choice of delivery technology depends on several groups of factors. One of those factors is the nature of the assessment material. If the material consists of a relatively simple stimulus material and multiple choice response options to be answered by clicking on a radio button (or even provision for a constructed text response) then the demands on the delivery technology will be relatively light. If the assessment includes rich graphical, video or audio material or involves students by using live software applications in an open authentic context then the demands on the delivery technology will be much greater. For the assessment of 21st century skills it is assumed that students would be expected to interact with relatively rich materials.

A second group of factors relates to the capacity of the connection of the school, or other assessment site, to the internet. There is considerable variation among countries, and even among schools within countries, in the availability and speed of internet connections in schools. In practice the capacity of the internet connection needs to provide for simultaneous connection of the specified number of students completing the assessment at the same time as other computer activity involving the internet is occurring. There are examples where the demand of concurrent activity (which may have peaks) has not been taken into account. In the 2008 cycle of the Australian national assessment of ICT literacy, which involved ten students working concurrently with moderate levels of graphical material and interactive live software tasks but not video, a minimum of 4 mbps was specified. In this project schools provided information about the computing resources and technical support that they had. They provided this information on a project website that uses the same technology as the preferred test-delivery system because the process of responding would provide information about internet connectivity (and the capacity to use that connectivity) and the specifications of the computer resources available. School internet connectivity has also proven to be difficult to monitor accurately. Speed and connectivity tests are only valid if they are conducted in the same context as the test-taking. In reality it is difficult to guarantee this equivalence, as the connectivity context depends both on factors within schools (such as concurrent internet and resource use across the school) and factors outside schools (such as competing

---

<sup>4</sup> The contributions of Julian Fraillon of ACER and Mike Janic of SoNET systems to these thoughts are acknowledged.

internet traffic from other locations). As a consequence it is necessary cautiously overestimate the necessary connection speed to guarantee successful internet assessment delivery. In the previously mentioned Australian national assessment of ICT literacy the minimum necessary standard of 4mbps per school was specified even though the assessment could run smoothly on a true connections speed of 1mbps.

A third group of factors relates to school computer resources. This includes having sufficient numbers of co-located computers and whether those computers are networked. If processing is to be conducted on local machines it includes questions of adequate memory and graphic capacity. Whether processing is remote or local aspects of screen size and screen resolution are important factors to be considered in determining an appropriate delivery technology. Depending on the software delivery solution being used it is also possible that school level software (in particular the type and version of the operating system and software plug-ins such as Java or ActiveX) can also influence the success of online assessment delivery.

### **Types of delivery technology**

There is a number of ways in which computer-based assessments can be delivered to schools. These can be classified in four main categories: those that involve delivery through the internet; those that involve through a local server connected to the school network; those that involve delivery on removable media; and those that involve delivery of mini-labs of computers to schools. The balance in the choice of delivery technology depends on a number of aspects of the IT context and changes over time as infrastructure improves, existing technologies develop and new technologies emerge.

#### *Internet-based delivery*

Internet access to a remote server (typically using an SSL-VPN internet connection to a central server farm) is often the preferred delivery method because the assessment software operates on a remote server (or server farm) and makes few demands on the resources of the school computers. Since the operation takes place on the server it provides a uniform assessment experience and enables student responses to be collected on the host server. This method solution minimizes, or even completely removes the need for any software installations on school computers or servers and eliminates the need for school technical support being involved in setting up and execution. It is possible to have the remote server accessed using a thin client that works from a USB stick without any installation to local workstations or servers.

This delivery method requires a sufficient number of co-located networked computers with access to an internet gateway at the school that has sufficient capacity for the students to interact with the material remotely without being compromised by other school internet activity. The bandwidth required will be dependent on the nature of the assessment material and the number of students accessing the material concurrently. In principle, where existing internet connections are not adequate, it would be possible to provide school access to the internet through the wireless (e.g. Next G) network but this is an expensive option for a large-scale assessment survey and is often least effective in remote areas where cable-based services are not adequate. In addition to requiring adequate bandwidth at the school, internet-based delivery is dependent on the bandwidth and capacity of the remote server to be able to accommodate multiple concurrent connections.

Security provisions installed on school and education system networks are also an issue for internet delivery of computer-based assessments. Those security provisions can block access to some ports and restrict access to non-approved internet sites. In general the connectivity of school internet connections is improving and is likely to continue to improve but security restrictions on school internet access seem likely to become stricter. It is also often true that responsibility for individual school-level security rests with a number of different agents. In cases where security is controlled at the school, sector and jurisdictional level the process of negotiating access for all

schools in a representative large-scale sample can be extremely time consuming, expensive and potentially eventually unsuccessful.

A variant of having software located on a server is to have an internet connection to a website but this usually means limiting the nature of the test materials to more static forms. Another variant is to make use of web-based applications (such as Google docs) but this involves limitations on the scope for adapting those applications and on control (and security) of the collection of student responses. An advantage is that it provides the applications in many languages. A disadvantage is that if there was insufficient bandwidth in a school it would not be possible to locate the application on a local server brought to the school. In principle it would be possible to provide temporary connections to the internet via the wireless network but at this stage it is expensive and not of sufficient capacity in remote area.

#### *Local server delivery*

Where internet delivery is not possible a computer-based assessment can be delivered on a laptop computer that has all components of the assessment software installed. This requires the laptop computer to be connected to the local area network (LAN) in the school and installed to function (by running a batch file) as a local server with the school computers functioning as terminals. When the assessment is complete the student response data can be delivered either manually (after being burned to CDs or memory sticks) or electronically (e.g. by uploading to a ftp site). The method requires a sufficient number of co-located networked computers one a laptop computer of moderate capacity to be brought to the school. This is a very effective delivery method that utilizes existing school computer resources but makes few demands on special arrangements.

#### *Delivery on removable media*

Early methods for delivering computer-based assessments to schools made use of compact disc (CD) technology. These methods of delivery limited the resources that could be included and involved complex provisions for capturing and delivering student response data. A variant that has been developed from experience of using laptop server technology is to deliver computer-based assessment software on Memory Sticks (USB or Thumb Drives) dispatched to schools by conventional means. The capacity of these devices is now such that the assessment software can work entirely from a Memory Stick on any computer with a USB interface. No software is installed on the local computer and the system can contain a database engine on the stick as well. This is self-contained environment that can be used to securely run the assessments and capture the student responses. Data can then be delivered either manually (e.g. by mailing the memory sticks) or electronically (e.g. by uploading data to an ftp site). After the data are extracted the devices can be re-used. The price is such that even regarding these as disposable is less than the cost of printing in a paper-based system. The method requires a sufficient number of co-located (but not necessarily networked) computers.

#### *Provision of mini-labs of computers*


For schools with insufficient co-located computers it is possible to deliver computer-based assessments by providing a set of student notebooks (to function as terminals) and a higher specification notebook to act as the server for those machines (MCEETYA, 2007). This set of equipment is called a mini-lab. The experience of this is that cable connection in the mini-lab is preferable to a wireless network because it is less prone to interference from other extraneous transmissions in some environments. It is also preferable to operate a mini-lab with a server laptop and clients for both cost consideration and for more effective data management. The assessment software is located on the "server" laptop and student responses are initially stored on that sever laptop. Data are transmitted to a central server either electronically when an internet connection is available or sent by mail on USB drives or CDs. Although this delivery method sounds expensive in a large project equipment costs have reduced substantially over recent years and amount to a

15 MINUTES

NAEP MATH ONLINE

QUESTION 2 OF 10 S1 G8

On the portion of the number line below, a dot shows where  $\frac{1}{2}$  is. Click on the number line to show where  $\frac{1}{8}$  is. If you need to change your answer, click on a different place on the number line.



PREVIOUS REVIEW NEXT

**Figure 5: Inserting a point on a number line**

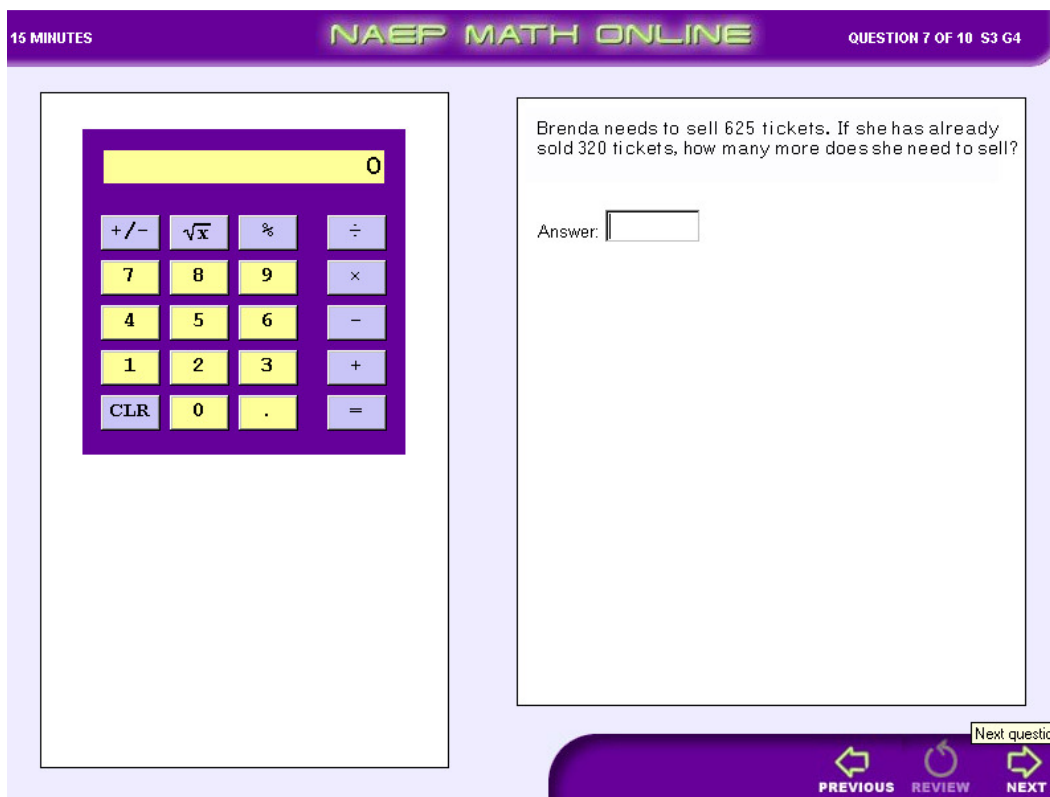
Source: Sandene, Bennett, Braswell & Oranje, 2005.

relatively small proportion of total costs. The difficulty with the method is managing logistics of delivering equipment to schools and moving that equipment from school to school as required.

### Use of delivery methods

All of these delivery technologies can provide a computer-based assessment that is experienced by the student in an identical way if the computer terminals at which the student works are similar. It is possible in a single study to utilize mixed delivery methods to make maximum use of the resources in each school. However, there are additional costs of development and licensing when multiple delivery methods are used. For any of the methods used in large-scale assessments (and especially those that are not internet based) it is preferable to have trained test administrators manage the assessment process or, at a minimum, to provide special training for school coordinators.

It was noted earlier in this section that the choice of delivery technology depends on the computing environment in schools and the optimum methods will change over time as infrastructure improves, existing technologies develop and new technologies emerge. In the Australian national assessment of ICT Literacy in 2005 (MCEETYA, 2007) computer based assessments were delivered by means of mini-labs of laptop computers (six per lab use in three sessions per day) transported to each of 520 schools. That ensured uniformity in delivery but involved a complex exercise in logistics. In the second cycle of the assessment in 2008 three delivery methods were used: internet connection to a remote server, a laptop connected as a local server on the school network and mini-labs of computers. The most commonly used method was the connection of a laptop to the school network as a local server being adopted in approximately 68 per cent of schools. Use of an internet connection to a remote server was adopted in 18 per cent of schools and the mini-lab method was adopted in approximately 14 per cent of the schools. The use of an internet connection to a remote server was more common in some education systems than others and in secondary compared to



**Figure 6: A numeric entry task allowing use of an onscreen calculator**

Source: Sandene, Bennett, Braswell, & Oranje, 2005.

primary schools (the highest being 34 per cent of the secondary schools in one State). Delivery by mini-lab was used in 20 per cent of primary schools and nine per cent of secondary schools. In the next cycle the balance of use of delivery technologies will change and some new methods (such as those based on memory sticks) will be available. Similarly the choice of delivery method will differ among countries and education systems depending on the infrastructure in schools, in education systems and more widely in those countries.

### Task presentation, response capture, and scoring

Technological delivery can be designed to closely mimic the task presentation and response entry characteristics of conventional paper testing. Close imitation is important if the goal is to create a technology-delivered test capable of producing scores comparable to a paper version. If, however, no such restriction exists, technological delivery can be used to dramatically change task presentation, response capture, and scoring.

#### Task presentation and response entry

Most technologically delivered tests administered operationally today use traditional item types; that is, questions that call for the static presentation of a test question and the entry of a limited response, typically a mouse click in response to one of a small set of multiple-choice options. In some instances, test questions in current operational tests call for more elaborate responses, such as the entry of an essay.

In between a multiple-choice response and an elaborate response format like an essay, lie a large number of possibilities and, as has been a theme throughout this paper, domain, purpose, and context play a role in how those possibilities are implemented and where they might work most

**Directions**

Decide if your answer is a

- Whole Number
- Decimal
- Fraction
- Mixed Number

Click on a box and type in the number. If you need to erase, use the Backspace Key.

Jorge left some numbers off the number line pictured above. Fill in the number that should go at **A**.

Pick one of the choices below. Type your number in.

Whole Number      Decimal      Fraction      Mixed Number

      .        /         /

PREVIOUS      REVIEW      NEXT

**Figure 7: A numeric entry task requiring use of a response template**

Source: Sandene, Bennett, Braswell, & Oranje, 2005.

appropriately. Below we give some examples for the three domain classes identified earlier: (1) domains in which practitioners interact with new technology primarily through the use of specialized tools; (2) domains in which technology may be used exclusively or not at all; and (3) domains in which technology use is central to the definition.

*Domains in which practitioners primarily use specialized tools*

As noted earlier, in mathematics, students and practitioners tend to use technology tools for specialized purposes, rather than pervasively in problem solving. Because such specialized tools as spreadsheets and graphing calculators are not used generally, the measurement of students' mathematical skills on computer has tended to track the manner of problem solving as it is conventionally practiced in classrooms and represented on paper tests, an approach which does not use the computer to maximum advantage. In this use, the computer serves primarily as a task presentation and response collection device and a key goal is preventing the computer from becoming an impediment to problem solving. That goal typically is achieved through both design and affording students the opportunity to become familiar with testing on computer and the task formats. Developing that familiarity might best be done through formative assessment contexts that are low stakes for all concerned.

The examples presented in figures illustrate the testing of mathematical competencies on computer that closely track the way those competencies are typically assessed on paper.

Figure 5 shows an item used experimentally in a study for the National Assessment of Educational Progress (NAEP) (Sandene, Bennett, Braswell, & Oranje, 2005). The task calls for the identification of a point on a number line that, on paper, would simply be marked by the student with a pencil. In this computer rendition, the student must use the mouse to click on the appropriate point on the line. Although this item format illustrates selecting from among choices, there is somewhat less of a



forced-choice flavor than the typical multiple-option item because there are many more points from which to choose.

In Figure 6 also from NAEP research, the examinee can use a calculator by clicking on the buttons, but then must enter a numeric answer in the response box. This process replicates what an examinee would do on a paper test using a physical calculator (i.e., compute the answer and then enter it onto the answer sheet). An alternative design for computer-based presentation would be to take the answer left in the calculator as the examinee's intended response to the problem.

An advantage to the use of an onscreen calculator is that the test developer controls when to make the calculator available to students (i.e., for all problems or for some subset). A second advantage is that the level of sophistication of the functions is also under the testing program's control. Finally, all examinees have access to the same functions and must negotiate the same layout. To ensure that all students are familiar with that layout, some amount of practice prior to testing is necessary.

Figure 7 illustrates an instance from NAEP research in which the computer appeared to be an impediment to problem solving. On paper the item would simply require the student to enter a value into an empty box represented by the point on the number line designated by the letter "A." Implementing this item on computer raised the problem of how to insure that fractional responses were input in the mathematically preferred "over/under" fashion, while not cueing the student to the fact that the answer was a mixed number. This response type, however, turned what was a one-step problem on paper into a two-step problem on computer because the student had to choose the appropriate template before entering the response. The computer version of the problem proved to be considerably more difficult than the paper version (Sandene, Bennett, Braswell, & Oranje, 2005).

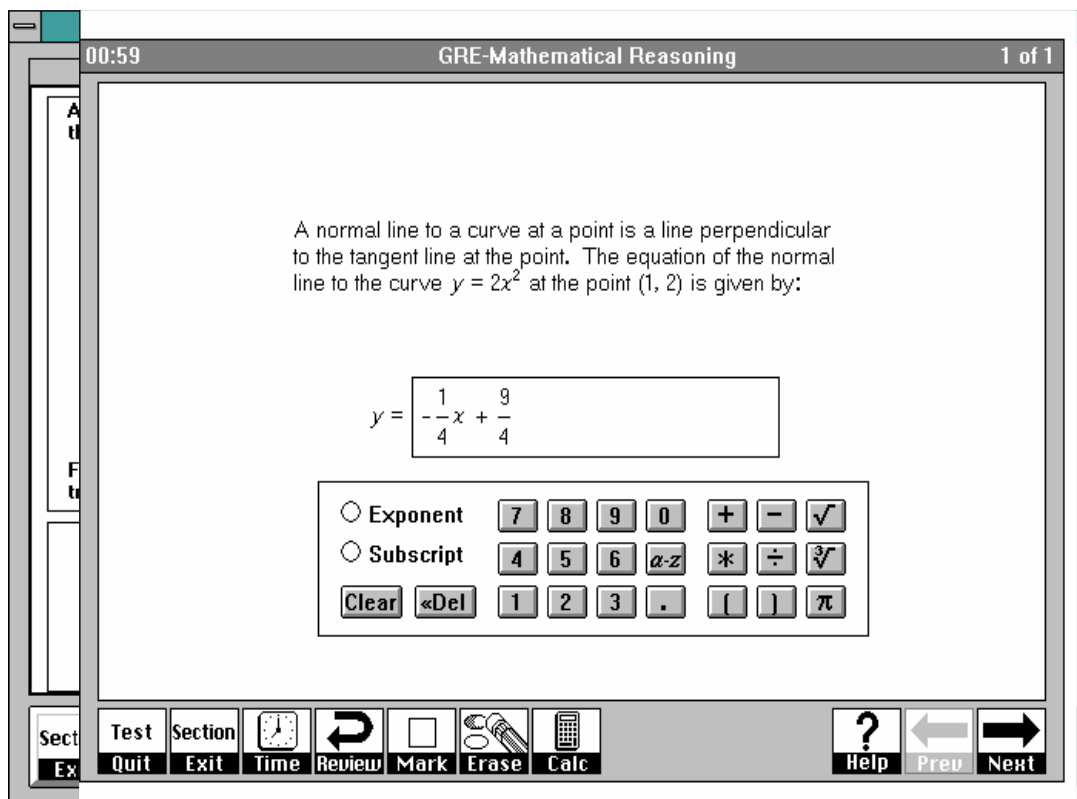


Figure 9: Task requiring symbolic expression for answer

Figure 8: Task with numeric answer that may not be scored automatically  
 Source: Bennett, Morley, & Quardt, 2000.

In Figure 8 is an example used in graduate admissions research (Bennett, Morley, & Quardt, 2000). Although only requiring the entry of numeric values, this response type is interesting for other reasons. The problem is cast in a business context. The stem gives three tables showing warehouses with inventory, stores with product needs, and the costs associated with shipping between warehouses and stores, as well as an overall shipping budget. The task is to allocate the needed inventory to each store (using the bottom table), without exceeding the resources of the warehouses or the shipping budget. The essence of this problem is *not* to find the best answer but only to find a reasonable one. Problems such as this one are typical of a large class of problems people encounter daily in real-world situations in which there are many right answers, the best answer may be too time consuming to find, and any of a large number of alternative solutions would be sufficient for many applied purposes.

One attraction of presenting this type of item on computer is that, even though there may be many correct answers, responses can be easily scored automatically. The scoring is done by testing each answer against the problem conditions. That is, does the student's answer fall within the resources of the warehouses, does it meet the stores' inventory needs, and does it satisfy the shipping budget? And, of course, many other problems with this same "constraint-satisfaction" character can be created, all of which can be automatically scored.

Figure 9 shows another type used in graduate admissions research (Bennett, Morley, & Quardt, 2000). The response type allows questions that have symbolic expressions as answers, allowing, for example, situations presented as text or graphics to be modeled algebraically. To enter an expression, the examinee uses the mouse to click on the onscreen keypad. Response entry is not as simple as writing an expression on paper. In contrast to the NAEP format above, this response type avoids the need for multiple templates while still representing the response in over/under

15 MINUTES

NAEP MATH ONLINE

QUESTION 10 OF 12 51 64

In which class can all the students be arranged in 4 rows with the same number of students in each row?  
Click on the class.

	Class 1	Class 2	Class 3
Number of Students	22	28	14

Explain your choice.

PREVIOUS REVIEW NEXT

**Figure 10: Task requiring forced choice and text justification of choice**

Source: Sandene, Bennett, Braswell & Oranje, 2005.

fashion. And, unlike paper, the responses can be automatically scored by testing whether the student's expression is algebraically equivalent to the test developer key.

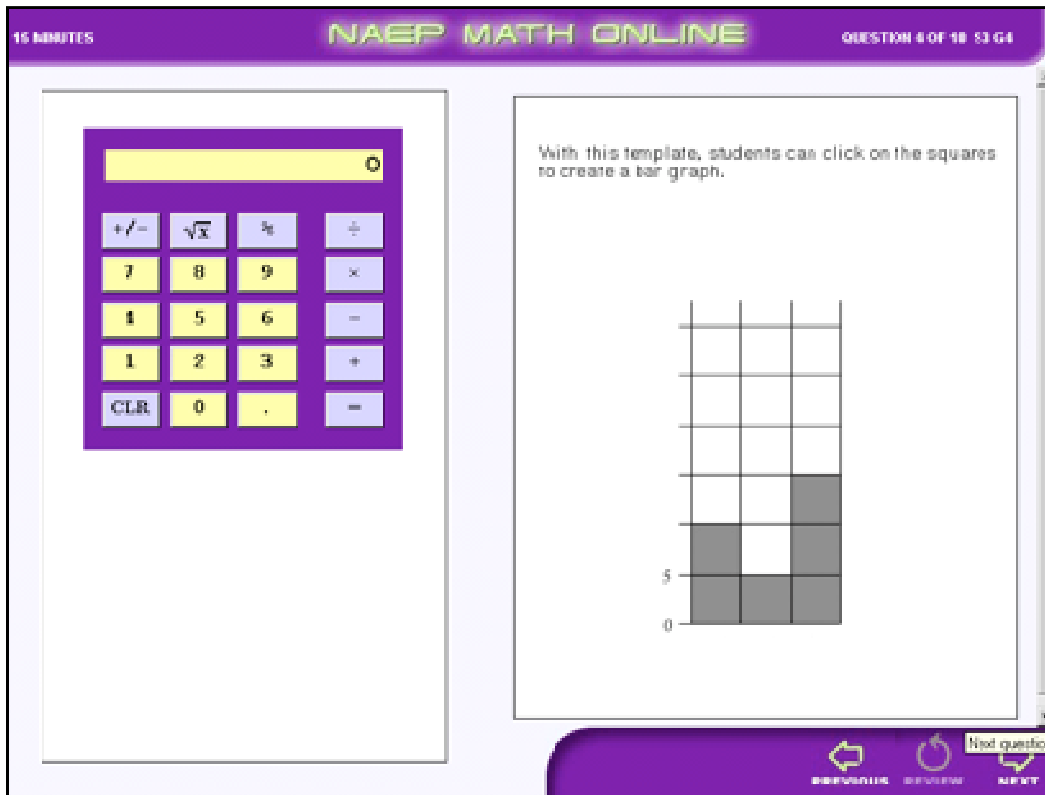
In Figure 10 is a question format from NAEP research in which the student must choose from among three options the class that has a number of students divisible by 4, and then enter text that justifies that answer. The written justification can be automatically scored but probably not as accurately as by human judges. Depending on the specific problem, the format might be used for gathering evidence related to whether a correct response indicates conceptual understanding or the level of critical thinking behind the answer choice.

Figure 11 shows a NAEP-research format in which the student is given data and then must use the mouse to create a bar graph representing those data. Bars are created by clicking on cells in the grid to shade or unshade a box.

Figure 12 shows a more sophisticated graphing task used in graduate admissions research. Here, the examinee plots points on a grid and then connects them by pressing a line or curve button. With this response type, problems can be presented that have one correct answer or multiple correct answers, all of which can be scored automatically. In this particular instance, a correct answer is any trapezoidal shape like the one depicted that shows the start of the bicycle ride at 0 miles and 0 minutes; a stop almost any time at 3 miles; and the conclusion at 0 miles and 60 minutes.

Finally, in the NAEP-research format shown in Figure 13, the student is asked to create a geometric shape, say a right triangle, by clicking on the broken line segments, which become dark and continuous as soon as they are selected. The advantage of this format over free-hand drawing, of course, is that the nature of the figure will be unambiguous and can be scored automatically.

In the response types above, the discussion has focused largely on the method of responding, as the stimulus display itself differed in only limited ways from what might have been delivered in a



**Figure 11: Graph construction with mouse clicks to shade/unshade boxes**

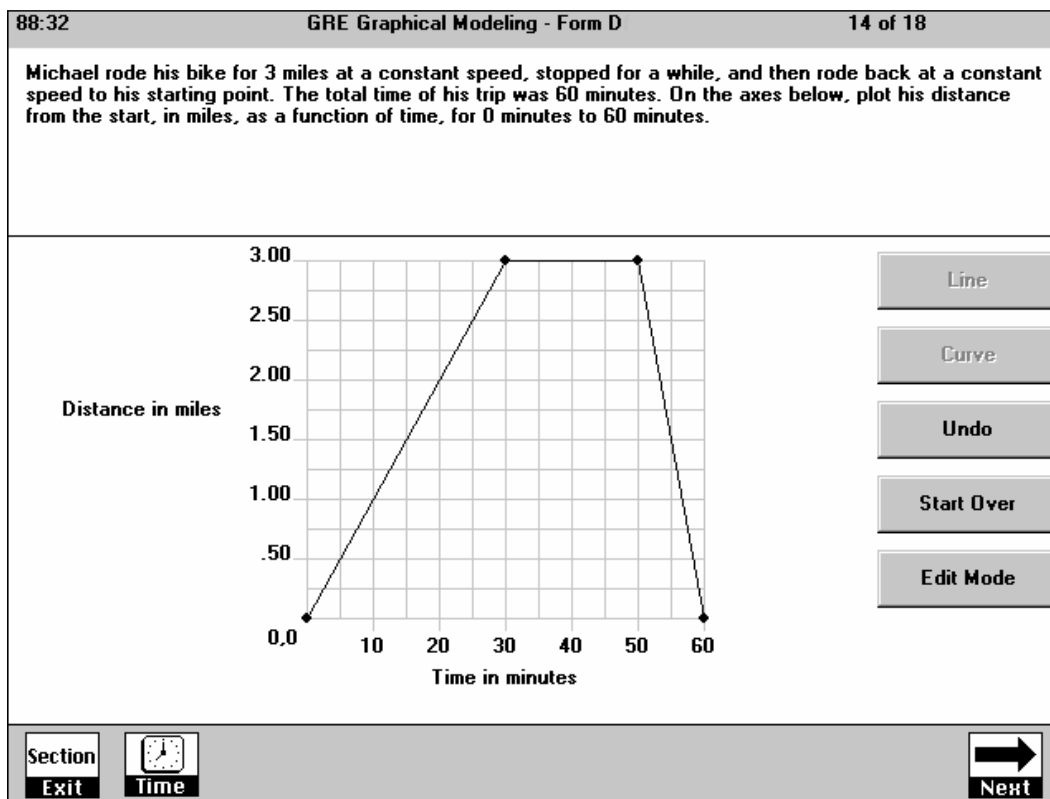
Source: Sandene, Bennett, Braswell, & Oranje, 2005.

paper test. And, indeed, the response types were generally modeled upon paper tests in an attempt to preserve comparability with problem solving in that format.

However, there are domains in which technology delivery can make the stimulus dynamic through the use of audio, video, or animation, an effect that cannot be achieved in conventional tests unless special equipment is used (e.g, TV monitor with video playback). Listening comprehension is one such domain where, as in mathematics, interactive technology is not used pervasively in schools as part of the typical domain practice. For assessment purposes, dynamic presentation can be paired with traditional test questions, as when a student is presented with an audio clip from a lecture and then asked to respond onscreen to a multiple-choice question about the lecture. Tests like the TOEFL iBT (Test of English as a Foreign Language Internet Based Test) pair such audio presentation with a still image, a choice that appears reasonable if the listening domain is intentionally conceptualized to exclude visual information. A more elaborate conception of the listening comprehension construct could be achieved if the use of visual cues is considered important by adding video of the speaker

Science is a third instance in which interactive technology is not used pervasively in schools as part of the typical domain practice. Here, again, interactive tools are used for specialized purposes, such as spreadsheet modeling or running simulations of complex physical systems. Response formats used in testing might include responding to forced-choice and constructed-response questions after running simulated experiments or after observing dynamic phenomena presented in audio, video, or animation.

There have been many notable projects that integrate the use of simulation and visualization tools to provide rich and authentic tasks for learning in science. Such learning environments facilitate a deeper understanding of complex relationships in many domains through interactive exploration

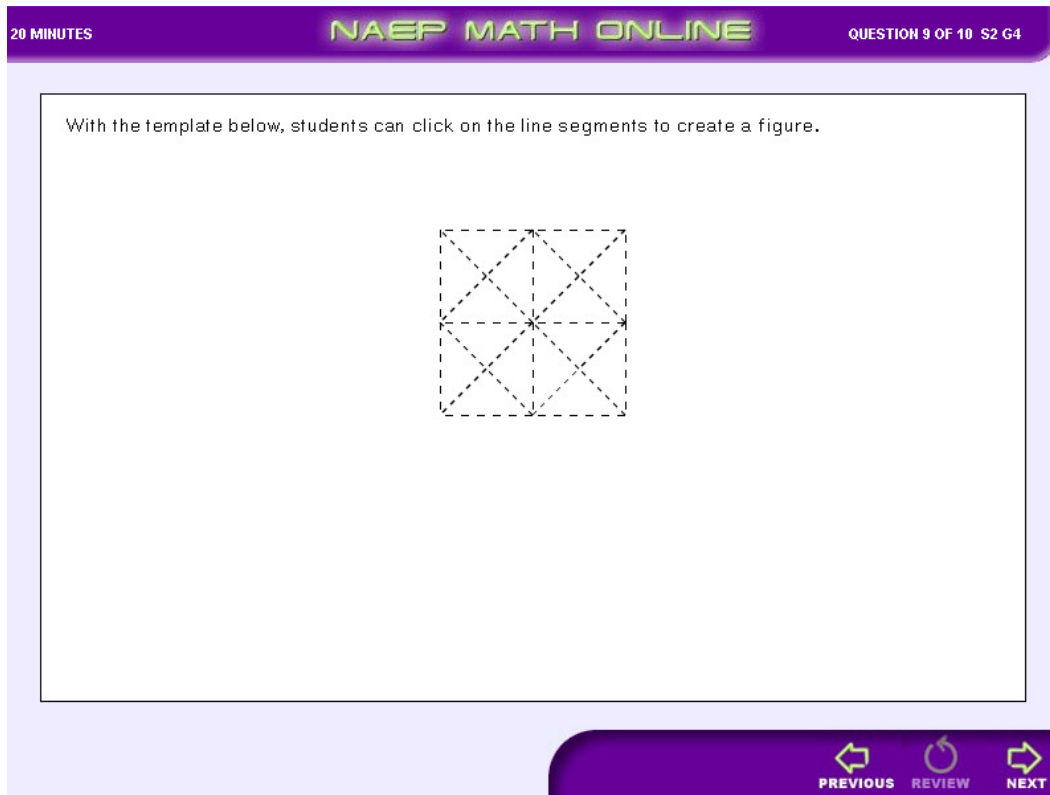


**Figure 12: Plotting points on grid to create a line or curve**

Source: Bennett, Morley, & Quardt, 2000.

(e.g. Mellar et al., 1994; Pea, 2002; Feurzeig and Roberts, 1999; Tinker and Xie, 2008). Many of the technologies used in innovative science curricula also have the potential to be used or adapted for use in assessment in science education that opens up new possibilities for what kinds of student performances can be examined for formative or summative purposes (Quellmalz and Haertel, 2004). Some examples of the integration of such tools in assessment in science is given below to illustrate the range of situations and designs that can be found in the literature.

Among the earliest examples of technology-supported performance assessment in science that targets non-traditional learning outcomes are the assessment tasks developed for the evaluation of the GLOBE environmental science education program. One of the examples described by Means and Haertel (2002) was designed to measure inquiry skills associated with the analysis and interpretation of climate data. Here, students were presented with a set of climate-related criteria for selecting a site for the next Winter Olympics as well as multiple types of climate data on a number of possible candidate cities. The students had to analyze the sets of climate data using the given criteria, decide on the most suitable site on the basis of those results and then prepare a persuasive presentation incorporating displays of comparative climatic data to illustrate the reasons for their selection. The assessment was able to reveal the extent to which students were able to understand the criteria and to apply them consistently and systematically, and whether they were able to present their argument in a clear and coherent manner. The assessment hence served the purpose of evaluating the GLOBE program well. However, Means and Haertel (2002) point out that as the assessment task was embedded within the learning system used in the program and cannot be used to satisfy broader assessment needs. One of the ways they have explored to overcome such limitations was the development and use of assessment templates to guide the design of classroom assessment tools.



**Figure 13: Item requiring construction of a geometric shape**

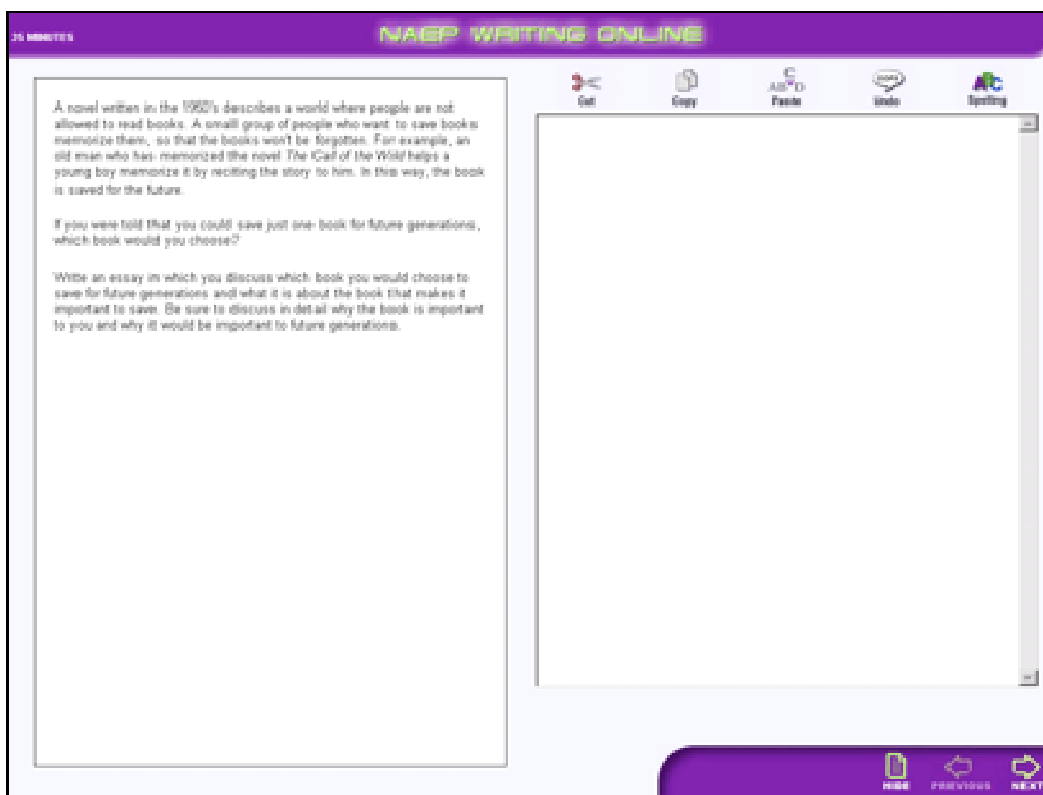
Source: Sandene, Bennett, Braswell, & Oranje, 2005.

The *SimsScientists* assessments is a project that makes use of interactive simulation technology for the assessment of students' science learning outcomes designed to support classroom formative assessment (Quellmalz and Pellegrino, 2009; Quellmalz, Timms and Buckley, 2009). The simulation-based assessments were designed according to an evidence-centered design model (Mislevy and Haertel, 2006) such that the task designed will be based on models that elicit evidence of the targeted content and inquiry targets defined in the student model, and the students' performance will be scored and reported based on an appropriate evidence model for reporting on students' progress and achievement on the targets. In developing assessment tasks for specific content and inquiry targets, much attention is given to the identification of major misconceptions related to the assessment targets reported in the science education research literature as the assessment tasks are designed to reveal incorrect or naïve understanding. The assessment tasks are designed as formative resources by providing: (1) immediate feedback according to the students' performance, (2) real-time graduated coaching support to the student and (3) diagnostic information that can be used for further offline guidance and extension activities.

*Domains in which technology may be used exclusively or not at all*

In the domain of writing, many individuals use the computer almost exclusively, while many other individuals use it rarely if at all. This situation has unique implications for design in that the needs of both types of individuals must be accommodated in writing assessment.

Figure 14 gives an example format from NAEP research. On the left is writing prompt and on the right is a response area that functions much like a simplified word processor. Six functions are available through tools above the response area, including cutting, copying and pasting text; undoing the last action; and checking spelling. Several of these functions are also accessible through standard keystroke combinations like Control-c for copying text.



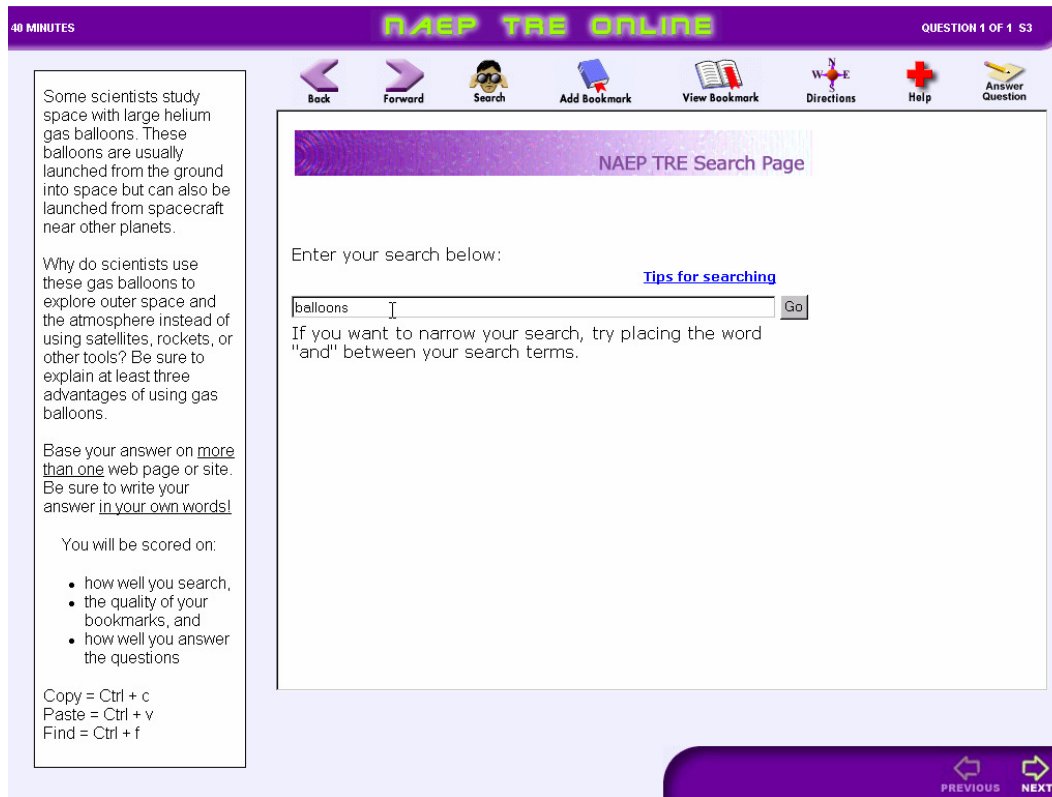
**Figure 14: A response type for essay writing**  
Source: Horkay, Bennett, Allen, Kaplan, & Yan, 2006.

This format was intended to be familiar enough in its design and features to allow those proficient in writing on computer to quickly learn it and easily use it almost as they would in their typical writing activity. All the same, the design may work to the disadvantage of students who routinely use the more sophisticated features of commercial word-processors.

The simple design of this response type was also intended to benefit the individual who doesn't write on computer at all. However, that individual would likely be disadvantaged by any design requiring keyboard input since computer familiarity, and particularly keyboarding skill, appears to affect online writing performance (Horkay, Bennett, Allen, Kaplan, & Yan, 2006). A more robust test design might also allow for handwritten input via a stylus. But even that input would require prior practice for those individuals not familiar with using a tablet computer. The essential point is that, for domains where some individuals practice primarily with technology tools and others do not, both technologies delivered and traditional forms of assessment may be necessary.

In writing assessment, as well as in other domains where a technology tool is employed, a key issue is whether to create a simplified version of the tool for use in the assessment or use the actual tool. Using the actual tool, in this instance, a particular commercial word processor, typically involves the substantial cost of licensing the technology (unless students use their own or institutional copies). That tool may also only run locally, making direct capture of response data by the testing agency more difficult. Third, if a particular word processor is chosen, that choice may advantage those students who use it routinely and disadvantage those who use a competitor product. Finally, process data may not be easy, or even possible, to capture.

At the same time, there are issues associated with creating a generic tool. Among those issues are deciding what features to include in its design, the substantial cost of and time needed for



**Figure 15: A simulated internet search problem**

Source: Bennett, Persky, Weiss, & Jenkins (2007).

development, and the fact that all students will need time to familiarized themselves with the resulting tool.

*Domains in which technology use is central to the definition*

Technology-based assessment can probably realize its potential most fully and rapidly in domains where the use of interactive technology is central to the domain definition. In such domains, neither the practice nor the assessment can be done meaningfully without the technology. Although it can be used in either of the other two domain classes described above, simulation is a key tool in this third class of domains because it can be used to replicate the essential features of a particular technology or technology environment within which to assess domain proficiency

An example can be found in the domain of electronic information search. Figure 15 shows a screen from a simulated Internet created for use in NAEP research (Bennett, Persky, Weiss, & Jenkins, 2007). On the left side of the screen is a problem statement, which asks the student to find out and explain why scientists sometimes use helium gas balloons for planetary atmospheric exploration. Below the problem statement is a summary of directions students have seen in more detail on previous screens. To the right, is a search browser. Above the browser are buttons for revisiting pages, bookmarking, going to the more extensive set of directions, getting hints, and going to a form to take notes or write an extended response to the question posed.

The database constructed to populate this simulated Internet consisted of some 5,000 pages taken from the real Internet, including pages devoted to relevant and irrelevant material. A simulated Internet was used to ensure standardization because, depending upon school technology policy and the time of any given test administration, different portions of the real could be available to students, and to prevent access to inappropriate sites from occurring under the auspices of NAEP.



**Figure 16: Environment for problem solving by conducting simulated experiments**

Source: Bennett, Persky, Weiss, & Jenkins (2007).

Each page in the database was rated for relevance to the question posed by one or more raters. To answer that question, students had to visit multiple pages in the database and synthesize their findings. Student performance was scored on both the quality of the answer written in response to the question and on the basis of search behavior. Among other things, the use of advanced search techniques like quotes of the not operator, the use of bookmarks, the relevance of the pages visited or bookmarked, and the number of searches required to produce a set of relevant hits factored into scoring.

Of particular note is that the exercise will unfold differently depending upon the actions the examinee takes—i.e., upon the number and content of search queries entered and the particular pages visited. In that sense, the problem will not be the same for all students.

A second example comes from the use of simulation for conducting experiments. In addition to the electronic information-search exercise shown above, Persky, Weiss, & Jenkins (2007) created an environment in which 8th grade students were asked to discover the relationships among various physical quantities by running simulated experiments. The experiments involved manipulating the payload mass carried by, and the amount of helium put into, a scientific gas balloon so as to determine the relationship of these variables with the altitude to which the balloon can rise in the atmosphere. The interface that the students worked with is shown in Figure 16.

Depending on the specific problem presented (see upper right corner), the environment allows the student to select values for the independent variable of choice (payload mass and/or amount of helium), make predictions about what will happen to the balloon, launch the balloon, make a table or a graph, and write an extended response to the problem. Students may go through the problem solving process in any order and may conduct as many experiments as desired. The behavior of the balloon is depicted dynamically in the flight window and on the instrument panel below, which gives

its altitude, volume, time to final altitude, payload mass carried, and amount of helium put into it. Student performance was scored on the basis of the accuracy and completeness of the written response to the problem and upon aspects of the process used in solution. Those aspects included whether the number of experiments and range of the independent variable covered were sufficient to discover the relationship of interest, whether tables or graphs were constructed that incorporated all variables pertinent to the problem, and whether experiments were controlled so that the effects of different independent variables could be isolated.

## Scoring

For multiple-choice questions, the scoring technology is well established. For constructed-response question types, including some of the ones illustrated above, the technology for machine scoring is only just emerging. Drasgow, Luecht, and Bennett (2006) describe three classes of automated scoring of constructed response.

The first class is defined by a simple match between the scoring key and the examinee response. The response type given in Figure 5, p.36 (requiring the selection of a point on a number line) would fall into this class, as would a reading passage that asks a student to click on the point at which a given sentence should be inserted, problems that call for ordering numerical values by dragging and dropping them into slots, extending a bar on a chart to represent a particular amount, or entering a numeric response. In general, responses like these can be scored objectively. For some of these instances, tolerances need to be set or the need for making fine distinctions in scoring precluded at the time of response entry. As an example, if a question directs the examinee to click on the point on the number line represented by 2.5, and the interface allows clicks to be made anywhere on the line, some degree of latitude in what constitutes a correct response will need to be permitted. Alternatively, the response type can be configured to accept only clicks at certain intervals.

A second problem class concerns what Drasgow et al. term static ones too complex to be graded by simple match. These problems are static in the sense that the task remains the same regardless of the actions taken by the student. Examples from this class include mathematical questions calling for the entry of expressions (Figure 8, p.40), points plotted on a coordinate plane (Figure 12, p.43), or numeric entries to questions having multiple correct answers (Figure 8). Other examples are problems requiring a short written response, a concept map, an essay, or a speech sample. Considerable work has been done on this category of automated scoring, especially for essays (Shermis & Burstein, 2003), and such scoring is used operationally for summative assessment purposes that have high stakes for individuals by several large testing programs including the Graduate Record Examinations (GRE) General Test, the Graduate Management Admission Test (GMAT), and the TOEFL iBT. The automated scoring of low-entropy speech (i.e., highly predictable speech) is also beginning to see use in summative testing applications and that for less predictable, high-entropy speech in low stakes, formative assessment contexts (Xi, Higgins, Zechner, & Williamson, 2008).

The third class of problems covers those instances in which the problem changes as a function of the actions the examinee takes in the course of solution. The electronic-search response type shown in Figure 15 falls into this class. These problems usually require significant time for examinees to complete and, due to their highly interactive nature, they produce extensive amounts of data; every keystroke, mouse click, and resulting event can be captured. Those facts suggest the need, and the opportunity, for using more than a correct end-result as evidence for overall proficiency and, further, for pulling out dimensions in addition to an overall proficiency. Achieving these goals, however, has proven exceedingly difficult since, invariably, only some of the reams of data produced may be relevant. What to capture and what to score should be based upon a careful analysis of the domain conceptualization and the claims one wishes to make about examinees, of the behaviors that would provide evidence for those claims, and of the tasks that will provide that evidence (Mislevy, Almond, & Lukas, 2004; Mislevy, Steinberg, Almond, & Lukas, 2006). Approaches to the scoring of problems in this class have been demonstrated for strategy use in scientific problem-solving (Stevens, Lopo, & Wang, 1996; Stevens & Casillas, 2006); problem

solving with technology (Bennett, Jenkins, Persky, & Weiss, 2003); patient management for medical licensure (Clyman, Melnick, & Clauser, 1995); and computer network troubleshooting (Williamson, Almond, Mislevy, & Levy, 2006).

For all three classes of constructed response, and for forced-choice questions too, computer delivery offers an additional piece of information not captured by a paper test. The additional dimension is timing. That information may involve only the simple latency of the response for multiple-choice questions and constructed response questions in the first class (i.e., simple match) described above, where the simple latency is the time between the item's first presentation and the examinee's response entry. The timing data will be more complex for the second and third problem classes. An essay response, for example, permits latency data to be computed within and between words, sentences, and paragraphs. Some of those latencies may have implications for measuring keyboard skills (e.g., within word), whereas others may be more suggestive of ideational fluency (e.g., between sentences).

The value of timing data will depend upon the assessment domain, purpose, and context. Among other things, timing information might be most appropriate for domains in which fluency and automaticity are critical (e.g., reading decoding, basic number facts), for formative assessment purposes (e.g., where some types of delay may suggest the need for skill improvement), and when the test has low stakes for students (e.g., for determining which students are taking the test seriously).

### **Validity issues raised by the use of technology for assessment**

Below we discuss several general validity issues and, within them, some of the implications for the use of technology for assessment in the three domain classes identified earlier: (1) domains in which practitioners interact with new technology primarily through the use of specialized tools; (2) domains in which technology may be used exclusively or not at all; and (3) domains in which technology use is central to the definition.

Chief among the threats to validity are the extent to which (1) an assessment fails to fully measure the construct of interest and (2) other constructs tangential to the one of interest inadvertently influence test performance (Messick, 1989). With respect to the first threat, no single response type can be expected to fully represent a complex construct, certainly not one as complex (and as yet undefined) as "21st century skills." Rather, each response type, and its method of scoring, should be evaluated theoretically and empirically as to the particular portion of the construct it represents. Ultimately, it is the complete measure itself, as an assembly of different response types, which needs to be subjected to evaluation regarding the extent to which it adequately represents the construct for some particular measurement purpose and context.

A particularly pertinent issue concerning construct representation and technology arises as a result of the advent of automated scoring (though it occurs in human scoring also). At a high level, automated scoring can be decomposed into three separable processes, feature extraction, feature evaluation, and feature accumulation (Drasgow, Luecht, & Bennett, 2006). Feature extraction involves isolating scorable components, feature evaluation entails judging those components, and feature accumulation consists of combining the judgments into a score or other characterization. In automated essay scoring, for example, a scorable component may be the discourse unit (e.g., introduction, body, conclusion), which are judged as present or absent, and then the number present combined with similar judgments from other scorable components (e.g., the average word complexity, average word length). The choice of what aspects of writing to score, how to judge those aspects, and how to combine the judgments all bring into play concerns for construct representation. Automated scoring programs, for example, tend to use features that are easily computable and to combine them in ways that best predict the scores awarded by human judges under operational conditions. Even when it predicts operational human scores reasonably well, such an approach may not provide the most effective representation of the writing construct (Bennett, 2006; Bennett & Bejar, 1998), omitting features that cannot be easily extracted from an

essay by machine and, among the features that are extracted, giving undue weight to those that human experts would not necessarily value very highly (Ben-Simon & Bennett, 2007).

The second threat, construct-irrelevant variance, also cannot be precisely identified without a clear definition of the construct of interest. Absent knowing the exact target of measurement, it can be difficult to identify factors that might be irrelevant. Here, too, an evaluation can be conducted at the level of the response type, as long as one can make some presumptions about what the test, overall, was *not* supposed to measure.

Construct under-representation and construct-irrelevant variance factor into a third consideration key to the measurement of domain classes 1 and 2, the comparability of scores between the conventional and technology-based forms of a test. Although different definitions exist, a common conceptualization is that scores may be considered comparable across two delivery modes when those modes produce highly similar rank orders of individuals and highly similar score distributions (APA, 1986, p. 18). If the rank-ordering criterion is met but the distributions are not the same, it may be possible to make scores interchangeable through equating. Differences in rank order, however, are usually not salvageable through statistical adjustment. A finding of score comparability between two testing modes implies that the modes represent the construct equally well and that neither mode is differentially affected by construct-irrelevant variance. That said, such a finding does not indicate that the modes represent the construct sufficiently for a given purpose, nor that they are uncontaminated by construct-irrelevant variance; it only implies that scores from the modes are equivalent in whatever it is that they measure. Last, a finding that scores are not comparable suggests that the modes differ in either their degree of construct representation, construct-irrelevant variance, or both.

Comparability of scores across testing modes is important when a test is offered in two modes concurrently and users wish scores from the modes to be interchangeable. Comparability may also be important when there is a transition from conventional to technology delivery and users wish to compare performance across time. There have been many studies of the comparability of paper and computer-based tests of cognitive skills for adults with the general finding that scores are interchangeable for power tests but not for speeded measures (Mead & Drasgow, 1993). In primary and secondary school populations, the situation is less certain (Drasgow, Luecht, & Bennett, 2006). Several meta-analyses have concluded that achievement tests produce comparable scores (Kingston, 2009; Wang, Jiao, Young, Brooks, & Olson, 2007, 2008). This conclusion, however, is best viewed as preliminary because the summarized effects have come largely from analyses of distribution differences without consideration of rank-order differences, multiple-choice measures, unrepresentative samples, non-random assignment to modes, unpublished studies, and from a few investigators without accounting for violations of independence. In studies using nationally representative samples of middle-school students with random assignment to modes, analyses more sensitive to rank order, and constructed-response items, the conclusion that scores are generally interchangeable across modes has not been supported (e.g., Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008; Horkay, Bennett, Allen, Kaplan, & Yan, 2006).

It should be evident that, for domain class 3, score comparability across modes plays no role because technology is central to the domain practice and, putatively, that practice cannot be measured effectively without using technology. For this domain class, only one testing mode should be offered. However, a set of claims about what the assessment is intended to measure and evidence about the extent to which those claims are supported is still essential, as it would be for any domain class. The claims and evidence needed to support validity take the form of an argument that includes theory, logic, and empirical data (Kane, 2006; Messick, 1989).

For domain class 1, where individuals interact with technology primarily through the use of specialized tools, assessment programs often choose to measure the entire domain on computer even though some (or even most) of the domain components are not typically practiced in a technology environment. This decision may be motivated by a desire for faster score turn-around or for other practical reasons. For those domain components not typically practiced on computer,

construct-irrelevant variance may be introduced into problem solving if the computer presentation used for assessment diverges too far from the typical domain (or classroom instructional) practice.

Figure 7 above illustrates such an instance from NAEP mathematics research in which the computer appeared to be an impediment to problem solving. In this problem, the student was asked to enter a value that represented a point on a number line. The computer version proved to be considerably more difficult than the paper version presumably because the former added a requirement not present in the paper mode (i.e., the need to select a response template before entering an answer) (Sandene, Bennett, Braswell, & Oranje, 2005). It is worth noting that this alleged source of irrelevant variance might have been trained away by sufficient practice with this response format in advance of the test. It is also worth noting that, under some circumstances, working with such a format might not be considered irrelevant at all (e.g., if such a template-selection procedure was typically used in mathematical problem solving in the target population of students).

Figure 9 above offers a second example. In this response type, created for use in graduate and professional admissions testing, the student enters complex expressions using a soft keypad (Bennett, Morley, & Quardt, 2000). Gallagher, Bennett, Cahalan, & Rock (2002) administered problems using this response type to college seniors and first-year graduate students in mathematics-related fields. The focus of the study was to identify whether construct-irrelevant variance was associated with the response-entry process. Examinees were given parallel paper and computer mathematical tests, along with a test of expression editing and entry skill. The study found no mean score differences between the modes, similar rank orderings across modes, and non-significant relations of each mode with the edit-entry test (implying that among the range of editing-skill levels observed, editing skill made no difference in mathematical test score). However, 77% of examinees indicated that they would prefer to take the test on paper were it to count, with only 7% preferring the computer version. Further, a substantial portion indicated having difficulty on the computer test with the response-entry procedure. The investigators then retrospectively sampled paper responses and tried to enter them on computer, finding that some paper responses proved too long to fit into the on-screen answer box. That finding suggested that some students might have tried to enter such expressions on the computer version but had to reformulate them to fit the required frame. If so, these students did their reformulations quickly enough to avoid a negative impact on their scores (which would have been detected by the statistical analysis). Even so, having to rethink and re-enter lengthy expressions likely caused unnecessary stress and time pressure. For individuals less skilled with computer than these mathematically adept college seniors and first-year graduate students, the potential for irrelevant variance would seem considerably greater.

In the design of tests for domain classes 1 and 2, there may be instances where comparability is *not* expected because the different domain competencies are intended to be measured across modes. For instance, in domain class 1, the conventional test may have been built to measure those domain components typically practiced on paper and the technology test built to tap primarily those domain components brought to bear when using specialized technology tools. In domain class 2, paper and computer versions of a test may be offered but, because those who practice the domain on paper may be unable to do so on computer (and vice versa), neither measurement of the same competencies nor comparable scores should be expected. This situation would appear to be the case in many countries among primary and secondary school students for purposes of summative writing assessment. Some students may be able to compose a timed response equally well in either mode but, as appeared to be the case for US 8<sup>th</sup> graders in NAEP research, many perform better in one or the other mode (Horkay, Bennett, Allen, Kaplan, & Yan, 2006). If student groups self-select to testing mode, differences in performance between the groups may become uninterpretable. Such differences could be the result of skill level (i.e., those who typically use one mode may be generally more skilled than those who typically use the other), mode (e.g., one mode may offer features that aid performance in ways that the other mode does not), or the interaction between the two (e.g., more skilled practitioners may benefit more from one mode than the other, while less skilled practitioners are affected equally by both modes).

An additional comparability issue relevant to computer-based tests *regardless* of domain class is the comparability of scores across hardware and software configurations, including between laptops and desktops, monitors of various sizes and resolutions, and screen-refresh latencies (as may occur due to differences in Internet bandwidth). There has been very little recent published research on this issue but the studies that have been conducted suggest that such differences can affect score comparability (Bridgeman, Lennon, & Jackenthal, 2001; Horkay, Bennett, Allen, Kaplan, & Yan, 2006). Bridgeman et al., for example, found reading comprehension scores to be higher for students taking a summative test on a larger, higher-resolution display than for students using a smaller, lower resolution screen. Horkay et al. found low-stakes summative test performance to be, in some cases, lower for students taking an essay test on a NAEP laptop than on their school computer, which usually was a desktop. Differences in, for example, keyboard and screen quality between desktops and laptops have greatly diminished over the past decade. However, the introduction of netbooks, with widely varying keyboards and displays, makes score comparability as a function of machine characteristics a continuing concern across domain classes.

Construct under-representation, construct-irrelevant variance, and score comparability all relate to the meaning or scores or other characterizations (e.g., diagnostic statements) coming from an assessment. Some assessment purposes and contexts bring into play claims that require substantiation beyond that related to the meaning of these scores or characterizations. Such claims are implicit, or more appropriately explicit, in the theory of action that underlies use of the assessment (Kane, 2006). A timely example is summative assessment such as that used under the US *No Child Left Behind Act*. Such summative assessment is intended not only to measure student (and group) standing, but explicitly to facilitate school improvement through various legally mandated, remedial actions. A second example is formative assessment in general. The claims underlying the use of such assessments are that they will cause greater achievement than would otherwise occur. In both the case of *NCLB* summative assessment and of formative assessment, evidence needs to be provided, first, to support the quality (i.e., validity, reliability, and fairness) of the characterizations of students (or institutions) coming from the measurement instrument (or process). Such evidence is needed regardless of whether those characterizations are scores or qualitative descriptions (e.g., a qualitative description in the summative case would be, “the student is proficient in reading; in the formative case, “the student misunderstands borrowing in two-digit subtraction and needs targeted instruction on that concept”). Second, evidence needs to be provided to support the claims of impact on individuals or institutions that the assessments are intended to have. Impact claims are the province of program evaluation and relate to whether use of the assessment had its intended effects on student learning, or on other classroom or institutional practices. It is important to realize that evidence of impact is required *in addition to*, not as substitute for, evidence of score meaning, even for formative assessment purposes. Both types of evidence are required to support the validity and efficacy arguments that underlie assessments intended to effect change on individuals or institutions (Bennett, 2009, p. 14-17; Kane, 2006, p. 53-56).

One implication of this separation of score meaning and efficacy is that assessments delivered in multiple modes may differ in score meaning, in impact, or in both. One could, for example, envision a formative assessment program offered on both paper and computer whose characterizations of student understanding and of how to adapt instruction were equivalent – i.e., equally valid, reliable, and fair – but that were differentially effective because the results of one were delivered faster than the results of the other.

### **Special applications and testing situations enabled by new technologies**

As it was already discussed in the previous sections, technology offers opportunities for assessment in domains and contexts where assessment otherwise would not be possible or would be difficult. Beyond extending the possibilities of routinely applied mainstream assessments, technology makes testing possible in several specific cases and situations. Two rapidly growing areas are discussed here; developments at both areas are driven by the needs of educational practices. Both areas of application still face several challenges, and exploiting the full potential of technology in these areas requires further research and developmental work.

## Assessing students with special educational needs

In modern societies, there are strong tendencies for teaching students, whose development, for whatever reason, is different from that of the typical one, together with their peers (mainstreaming, inclusion, integration etc.). Furthermore, the ones, who face challenges, are provided with extra care and facilities to overcome the difficulties (the principle of equal educational opportunities). Students, who need this type of special care, will be called students with Special Educational Needs (SEN) here. The definition that classifies SEN students broadly change from country to country, therefore, the proportion of SEN students within a population may vary in a broad range. Taking all kinds of special needs into account, in some countries their proportion may be up to 30%. This number indicates that using technology to assess SEN students is not a marginal issue, and using technology may vitally improve many students' chance for success in education, and later for leading a complete life.

The availability of specially trained teachers and experts often limits the fulfillment of these educational ideals, but technology may often fill the gaps. In several cases, using technology instead of relying on the services of human helpers is not only a replacement with limitations, but an enhancement of the personal capabilities of SEN students, that makes independent learning possible.

In some cases, there may be a continuum between slow (but steady) development, temporal difficulties, and specific developmental disorders. In other cases, development is severely hindered by factors, early identification and treatment of which may help to solve the problems. In the most severe cases, personal handicaps cannot be altered, and technology is used to improve functionality.

As the inclusion of students with special educational needs in regular classrooms is an accepted basic rule, there is a growing demand to assess those students together, who are taught together (see chapter 12 of Koretz, 2008). Technology may be applied in this process in a number of different ways.

- Scalable fonts, using larger fonts.
- Speech synthesizers for reading texts.
- Blind students may enter responses on specific keywords.
- Development of a large number of specific technology-based diagnostic tests is in progress. TBA may reduce the need for specially trained experts, and improve the precision of measurement, especially in the psychomotor area.
- Customized interfaces devised for physically handicapped students. From simple instruments to sophisticated eye tracking make testing accessible for students with a broad range of physical handicaps (Lőrincz, 2008).
- Adapting tests to the individual needs of students. The concept of adapting testing may be generalized to identify some types of learning difficulties and to offer items matched to students' specific needs.
- Assessments built in specific technology-supported learning programs. A reading improvement and speech therapy program recognizes the intonation, the tempo and the loudness of speech (or loud reading) and compares to the pre-recorded standards, and provides visual feedback to students (<http://www.inf.u-szeged.hu/beszedmester>).

At present, the technologies are already available, and many of them are routinely used in e-learning (Ball, S. et al., 2006; Reich & Petter, 2009). However, transferring and implementing these technologies into the area of TBA requires further developmental work. Including SEN students in mainstream TBA assessment is desired on the one hand, but expressing their achievements on the same scale raises several methodological and theoretical issues.

## **Connecting individuals: assessing collaborative skills and group achievement**

Sfard (1998) distinguishes two main metaphors in learning: learning as acquisition and learning as participation. CSCL and collaborative learning in general belong more to the participation metaphor, which focuses on learning as becoming a participant, and interactions through discourse and activity as the key processes. Depending on the theory of learning underpinning the focus on collaboration, the learning outcome to be assessed may be different (Dillenbourg, 1996). Assessing learning as an individual outcome is consistent with a socio-constructivist or socio-cultural view of learning as social interaction provides conditions that are conducive to conflict resolution in learning (socio-constructivist) or scaffold learning through bridging the zone of proximal development (socio-cultural). On the other hand, a shared cognition approach to collaborative learning (Suchman, 1987; Lave, 1988) considers the learning context and environment an integral part of the cognitive activity and a collaborating group can be seen as forming a single cognizing unit (Dillenbourg, 1996) and assessing learning beyond the individual poses an even bigger challenge.

Webb (1995) provides an in-depth discussion of the theoretical and practical challenges of assessing collaboration in large-scale assessment programs based on a comprehensive review of studies on collaboration and learning. In particular, she highlights the importance of defining clearly the purpose of the assessment, and giving serious consideration to the goal of group work and the group processes that are supposed to contribute to those goals to ensure that these work toward rather than against, the purpose of the assessment. Three purposes of assessment were delineated in which collaboration plays an important part: the level of an individual's performance after learning through collaboration, group productivity and an individual's ability to interact and function effectively as a member of a team. Different assessment purposes entail different group tasks. Group processes leading to good performance is often different depending on the task and could be competing. For example, if the goal of the collaboration is on group productivity, giving time to explain to each other to enhance individual learning through collaboration may lower group productivity within a given period of time. The purpose of the assessment should also be made clear, as this will influence individual behavior in the group. If the purpose is to measure individual student learning, Webb suggests that the test instructions should focus on individual accountability and individual performance in the group work, and to include in the instruction what constitutes desirable group processes and why. On the other hand, a focus on group productivity may act against equality of participation and may even lead to a socio-dynamic in which low-status members' contributions are ignored. Webb's paper also reviewed studies on group composition (in terms of gender, personality, ability, etc.) and group productivity. The review clearly indicates that group composition is one important issue in large-scale assessments of collaboration.

Owing to the complexities in assessing cognitive outcomes in collaboration, global measures of participation such as frequency of response or the absence of disruptive are often used as indicators of collaboration, which falls far from being able to reveal the much more nuanced learning outcomes such as being able to explore a problem, generate a plan or design a product. Means, Penuel and Quellmalz (2000) describe a Palm-top Collaboration Assessment project in which they developed an assessment tool that teachers can use for "mobile real-time assessments" of collaboration skills as they move among groups of collaborating students. Teachers can use the tool to rate each group's performance on nine dimensions of collaboration (p.9):

- Analyzing the Task
- Developing Social Norms
- Assigning and Adapting Roles
- Explaining/Forming Arguments
- Sharing Resources
- Asking Questions
- Transforming Participation
- Developing Shared Ideas and Understandings
- Presenting Findings



Teachers' ratings would be made on a three-point scale for each dimension, which would be stored on the computer for subsequent review and processing.

Unfortunately, research that develops assessment tools and instruments independent of specific collaboration contexts such as the above is rare, even though studies of collaboration and CSCL are becoming an important area in educational research. On the other hand, much of the literature on assessing collaboration, whether computers are being used or not, are linked to research on collaborative learning contexts. These may be embedded as an integral part of the pedagogical design such as in peer- and self-assessment (e.g. Boud, Cohen and Sampson, 1999; McConnell, 2002; Macdonald, 2003), and the primary aim is to promote learning through collaboration. The focus of some studies involving assessment of collaboration is on the evaluation of specific pedagogical design principles. Lehtinen et al. (1999) summarizes the questions addressed in this kind of studies as belonging to three different paradigms. "Is collaborative learning more efficient than learning alone?" is typical of questions under the effects paradigm. Research within the conditions paradigm studies how learning outcomes are influenced by various conditions of collaboration such as group composition, task design, collaboration context and the communication/collaboration environment. There are also studies that examine group collaboration development in terms of stages of inquiry (e.g. Gunawardena, Lowe and Anderson, 1997), demonstration of critical thinking skills (e.g. Henri, 1992) and stages in the development of a socio-metacognitive dynamic for knowledge building within groups engaging in collaborative inquiry (e.g. Law, 2005).

In summary, in assessing collaboration, both the unit of assessment (individual or group) and the nature of the assessment goal (cognitive, metacognitive, social or task productivity) can be very different. This poses serious methodological challenges to what and how this is to be assessed. Technological considerations and design are subservient to these more holistic aspects in assessment.

### **Need for further research and development**

In this section, first we present some general issues and directions for further research and development. Three main topics will be discussed which are more closely related to the technological aspects of assessment and add further subjects to the ones elaborated in the previous parts of this paper. Finally, a list of research themes will be presented that can be formed into research projects in the near future. These themes are more closely associated with the issues elaborated in the previous sections and focus on specific problems.

### **Migration strategies**

Compared to other educational computer technologies, Computer-Based Assessment bears additional constraints related to the measurement quality, as already discussed before. If the use of new technologies is sought to widen the range of skills and competencies one can address, or to improve the instrument in various aspects, special care should be taken when increasing the technological complexity or the user experience richness while maintaining the objective of an unbiased high quality measurement. Looking at new opportunities offered by novel advanced technologies, one can follow two different approaches: either considering technological opportunities as a generator of assessment opportunities, or carefully analyzing assessment needs to derive technological requirements that are mapped with available solutions or translated in new solution designs. At first sight, the former approach sounds more innovative than the later, which sounds more classical. However, both bear advantages and disadvantages that must be mitigated with the assessment context and the associated risk tolerance. The "technology opportunistic" approach has major inherent strength that had already been discussed in this paper by offering a wide range of new potential instruments providing a complete assessment landscape. Besides this strength, it potentially opens the door to new time and cost effective measurable dimensions that were never thought of before. As a drawback, it currently requires tremendous needs of long and costly validations. Underestimating this will certainly lead to the uncontrolled use and proliferation of

appealing but invalid assessment instruments. The later approach is not neutral either. While appearing more conservative and probably more suitable in mid- and high-stake contexts as well as in systemic studies, it also bears inherent drawbacks. Indeed, even if it guarantees the production of well-controlled instruments and measurement setting developments, it may also lead to mid- and long-term time consuming and costly operations that may hinder innovation by thinking “in the box”. Away from the platform approach, it may bring value by its capacity to address very complex assessment problems with dedicated solutions with the risk of discrepancies between actual technology literacy of the target population and “old-fashion” assessments diminishing the subject engagement. In other word and to paraphrase the US Web-Based Education Commission (cited in Bennett, 2001), measuring today’s skills with yesterday’s technology.

In mid- and high-stake individual assessments or systemic studies, willing to accommodate innovation while maintaining the trend at no extra cost (in terms of production as well as logistics) may seem illusive at first sight. Certainly, in these assessment contexts, unless a totally new dimension or domain is defined, disruptive innovation would probably never appear, and may not be sought at all. There is however a strong opportunity for academic interest to perform ambitious validation studies using frameworks and instruments built on new technologies. Taking into account the growing intricacy of psychometric and IT issues, no doubt that the most successful studies will be strongly inter-disciplinary. The intertwine between computer delivery issues in terms of cost together with software/hardware universality and the maintenance of trends and comparability represents the major rationale that calls for inter-disciplinarity.

### **Security, availability, accessibility, comparability**

Security is of utmost importance in high-stake testing. In addition to assessment reliability and credibility, security issues may also strongly affect the business of major actors in the fields. Security issues in Computer-Based Assessment depend on the purposes and contexts of assessments, on processes, and include a large range of issues.

The International Standard Institute has published a series of normative texts covering the Information security known as the ISO 27000 family. Among these standards, ISO 27001 specifies requirements for information security management systems, ISO 27002 describes a code of Practice for Information Security Management, and ISO 27005 covers the topic of information security risk management.

In the ISO 27000 family, Information security is defined according to three major aspects: the preservation of *confidentiality* (ensuring that information is accessible only to those authorized to have access), the preservation of information *integrity* (guaranteeing the accuracy and completeness of information and processing methods) and the preservation of information *availability* (ensuring that authorized users have access to information and associated assets when required). Security issues covered by the standards are of course not restricted to technical aspects. They also consider organizational and more social aspects of security management. For instance, abandoning a copy of an assessment on someone’s desk induces risks at the level of confidentiality and may be at the level of availability. Social engineering is also another example of non-technical security thread for password protection. These aspects are of equal importance in both paper-and-pencil and Computer-Based Assessment.

The control of test-taker identity is classically made using various flavors of login/ID and password protection. This can be complemented by additional physical ID verification. Proctoring techniques have also been implemented to enable the test-takers to start the assessment after having checked if the right person is actually taking the test. Technical solutions making use of biometric identification may help reducing the risks associated to identity. As a complementary tool, the generalization of electronic passports and electronic signatures should also be considered as a potential contribution to the improvement of identity control.

Classically, in high-stake assessment, the test administrator is in charge of detecting and preventing cheating when the test is administered centrally. A strict control of the subject with respect to assessment rules before the assessment takes place is a minimal requirement. Besides the control, a classical approach to prevent cheating is the randomization of items or the delivery of different sets of booklets with equal and proven difficulty. The later solution should be systematically selected because randomization of items poses other fairness problems that might disadvantage or advantage some test-takers (Marks & Cronje, 2008). In addition to test administrator control, cheating detection can be made by analyzing the behavior of the subject during the test administration. Computer forensic principles have been applied to the computer-based assessment environment to detect infringement to assessment rules. The experiment showed that typical infringement such as illegal communication making use of technology, use of forbidden software or devices, falsifying its identity, or getting access to material of another student can be detected using logs of all computer actions (Laubscher *et al.*, 2005).

Secrecy, availability and integrity of computerized tests and items, of personal data (to ensure privacy), and of the results (to prevent loss, corruption or falsifications) are usually ensured by classical IT solutions such as firewalls at server level, encryptions, certificates and strict password policy at both server, client and communication network levels, together with tailor organizational procedures.

Brain dumping is a severe issue that has currently not be circumvented satisfactorily in high-stake testing. Brain dumping is a fraudulent practice consisting in participating to a high-stake assessment session (paper-based or computer-based) in order to memorize a significant number of items. When organized at a sufficiently large scale with many fake test-takers, one is able to reconstitute an entire item bank. After having solved the items with domain experts, the item bank can be disclosed on the Internet, or sold to assessment candidates. More pragmatically and in a more straightforward way, an entire item bank can also be stolen and further disclosed by simply shooting pictures of the screens using a mobile phone camera or miniaturized web cams. From a research point of view, as well as from a business value point of view, this very challenging topic should be paid more attention from the research community. In centralized high-stake testing, potential tracks to address the brain dump problem and the screenshot problem are twofold. On one hand, one can evaluate technologies to monitor the test-taker activity on and around the computer and elaborate alert patterns, and on the other hand, one can design, implement, and experiment technological solutions at software and hardware levels to prevent test-taker to take pictures of the screen.

Availability of tests and item during the whole assessment period is also a crucial issue. In the case of Internet-based testing, different risks may be identified, such as highjacking of the web site or denial of service attacks, among others. Added to risks associated to cheating in general, Internet is not yet suitable for high- or mid-stake assessment. However, solutions might be found to make the required assessment and related technology available everywhere (ubiquitous), every time necessary, while overcoming the technological divide.

Finally, we expect that, from a research and development perspective, the topic of security in high-stake testing will be envisioned in a more global and multidimensional way, incorporating in a consistent solution framework all the aspects that have been briefly described here.

### **Ensuring framework and instrument compliance with model-driven design**

Current assessment frameworks tend to describe a subject area along two dimensions – the topics to be included and a range of actions that drive item difficulty. However, the frameworks do not necessarily include descriptions of the processes subjects use in responding to the items. Measuring these processes depends on more fully described models that can then be used not only to develop the items, or set of items associated with a simulation, but also to determine the functionalities needed in the computer-based platform. The objective is to enable a direct link between the conceptual framework of competencies to be assessed and the structure and

functionalities of the item type or template. Powerful modeling capacities can be exploited for that purpose. It would enable one to:

- maintain the semantics of all item elements and interactions and to guarantee that any of these elements is directly associated with a concept specified in the framework;
- maintain the consistency of the scoring along all sets of items (considering automatic, semi-automatic, or human scoring);
- help ensuring that what is measured is indeed what one intends to measure; and
- significantly enrich the results for advanced analysis by linking through complete traceability the performance/ability measurement, the behavioral/temporal data, and the assessment framework.

It is however important to note that while IT can offer a wide range of rich interactions that might assess more complex or more realistic situations, IT may also lead to other important biases if not properly grounded on a firm conceptual basis. Indeed, offering respondents interaction patterns and stimulations that are not part of a desired conceptual framework, may introduce performance variables that do not pertain to the measured dimension. As a consequence realism and attractiveness which may add to motivational and playability might introduce unwanted contributions into the measurement instead of enriching or improving the instrument. To exploit the capabilities offered by IT to build complex and rich items and tests to better assess competencies in various domains, one must be able to maintain a stable, consistent, and reproducible set of instruments. If full traceability between the framework and each instrument is not strictly maintained, the risk of mismatch becomes significantly higher. This would undermine the instrument validity and in turn the measurement validity. In a general sense the chain of decision traceability in assessment design covers a important series of steps from the definition of the construct, skill, domain, or competency to the final refinement of computerized items and tests, through the design of the framework, the design of items, the item implementation, the item production. At each step, the design and implementation would most probably gain quality if it refers to a clear and well-formed meta-model while in the meantime systematically referring to pieces from the previous steps.

This claim is at the heart of the Model-Driven Architecture (MDA) software design methodology proposed by the Object Management Group (OMG). Quality and interoperability arise from the independence of the system specification with respect to the system implementation technology. The final system implementation in a given technology results from formal mappings of system design to many possible platform (Poole, 2001). In OMG's vision, MDA enables improving maintainability of software (and consequently decrease costs and reduce delays) among other benefits breaking the myth of standalone application that require never ending corrective and evolutive maintenance (Miller & Mukerji, 2003).

In a more general fashion, the approach relates to Model-Driven Engineering relying on a series of components. Domain specific modeling languages (DSL) are formalized using meta-models, which define the semantics and constraints of concepts pertaining to a domain and their relationships. These DSL components are used by designers to express their design intention declaratively, within a close, common and explicit semantics, as instances of the meta-model (Schmidt, 2006). As many meta-models than actual facets of the domain requires can be used to embrace the complexity and to address specific aspects of the design using the semantics, paradigms, and vocabulary of the different experts specialized in each individual facet. The second fundamental component consists in transformation rules, engines and generators used to translate the conceptual declarative design into other model closer to the executable system. This transformational pathway from the design to the executable system can include more than one step depending on the number of aspects of the domain, together with the operational and organizational production processes. In addition to the above-mentioned advantages in terms of interoperability, system evolution and maintenance, from a pure conceptual design point of view, this separation of concerns has several advantages: First, it keeps the complexity at a manageable level; second, it segments design activities centers on each specialist field of expertise; third, it enables a full traceability of design decisions.

The later advantage is at the heart of design and final implementation quality and risk mitigation. As an example, these principles have been successfully applied in the fields of Business Process engineering to derive business processes and e-business transactions through model chaining deriving economically meaningful business processes from value models obtained by transforming an initial business model (Bergholtz *et al.*, 2005), (Schmitt & Grégoire, 2006). In the field of Information System engineering, Turki *et al.* have proposed an ontology-based framework to design an Information System through a stack of models addressing different abstractions of the problem as well as various facets of the domain, including legal constraints. Applying a MDE approach, their framework consists in a *conceptual map to represent ontologies and a set of mapping guidelines from conceptual maps into other object specification formalisms*. (Turki, Aïdonis, Khadraoui & Léonard, 2004). A similar approach has been used to transform natural language mathematical documents into computerized narrative structure that can be further manipulated (Kamareddine, Lamar, Maarek & Wells, 2007). The transformation relies on a chain of model instantiations addressing different aspects of the document including syntax, semantics, and rhetoric (Kamareddine, Maarek, Retel & Wells, 2007), (Kamareddine, Lamar, Maarek & Wells, 2007).

The hypothesis and expectation is that such a design approach will ensure the compliance between assessment intentions and the data collection instrument. Compliance is to be understood here as the ability to maintain the links between originating design concepts articulated according to the different facets of the problem and derived artifacts (solutions), along all the steps of the design and production process. Optimizing the production process, reducing the cost by relying on (semi-) automatic model transformation between each steps, enabling conceptual comparability of instrument, and possibly measuring their equivalence or divergence, and finally guaranteeing better data quality with reduced bias, are among the other salient expected benefits.

The claim for a platform approach independent from the content, based on a knowledge modeling paradigm, i.e., including ontology-based metadata management, has a direct relationship in term of solution opportunities to tackle the formal design and compliance challenge. Together with web technologies enabling distant collaborative work through the Internet, one can envision a strong promising answer to the challenges.

To set up a new assessment design framework according to the MDE approach, several steps should be fulfilled that require intensive research and development work. First, one has to identify the various facets and domain expertise that are involved in assessment design and organize them as an assessment design process. This step is probably the easiest one, and mostly requires a formalization effort. The more conceptual spaces bears inherent challenges of capturing the knowledge and expertise of experts in an abstract way to build the reference meta-models and their abstract relationships that will serve as a basis to construct the specific model instances pertaining to each given assessment on all important aspects. Once these models are obtained, the dedicated instrument design and production chain is set up and the process can be started. The resulting instances of this layer will consist in a particular construct, framework, and item, depending on the facet being considered. Validation strategies are still to be defined, as well support tool design.

The main success factor of the operation fundamentally resides in inter-disciplinarity. Indeed, to reach the adequate level of formalism and in order to provide the adequate IT support tools to designers, assessment experts should work in close connection with computer-based assessment and IT experts who can bring their well established arsenal of more formal modeling techniques. It is expected that this approach will improve the measurement quality by providing more formal definitions of the conceptual chain linking the construct concepts to the final computerized instrument. Thus minimizing the presence of item features or content that bears little or no relationship to the construct. When looking at the framework facet, the identification of indicators and their relationships, the quantifiers (along with their associated quantities) and qualifiers (along with their associated classes), and data receptors enabling the collection of information used to value or qualify the indicators must be unambiguously related to both construct definition and item interaction patterns, as well as providing explicit and sound guidelines for item designers with regard to scenario and item characteristic descriptions. Similarly, the framework design also serves

as a foundation from which to derive exhaustive and unambiguous requirements for the software adaptation of extension, from the item interaction and item runtime software behavior perspective. As a next step, depending on the particular assessment characteristics, item developers will enrich the design by instantiating of the framework in the form of a semantically embedded scenario. The scenario includes the definition of stimulus material, tasks to be completed, and response collection modes. Dynamic aspects of the items may also designed in the form of storyboards. Taking into account the scoring rules defined in the framework, expected response patterns are defined. As a possible following step, IT specialist will translate the item design into a machine-readable item description format. This constitutes the transposition of the item from a conceptual design to a formal description of the design in computer form, i.e., transforming a *descriptive* version to an *executable* or *rendered* version. Following the integrative hypermedia approach, the models involved in this transformation are the different media models and the integrative model.

### **Potential themes for research projects**

In this section, a list of research themes is presented. The themes listed here are not elaborated in details yet. Some of them are closely related and are highlighting different aspects of the same issue. These questions may later be grouped and organized into larger themes depending on the time frame, size and complexity of the proposed research projects. Several topics proposed here may be combined with the themes proposed by other working groups to form larger research projects.

1. Making sense of the hundreds of pieces of information students may produce when engaging in a complex assessment, such as a simulation. How to determine which actions are meaningful, and how to combine those pieces into evidence of proficiency, is an area that needs concentrated research. The work on evidence centered design by Mislevy and colleagues represents one promising approach to the problem.
2. Included in the above but probably requiring special mention is the issue of response latency. In some tasks and contexts, timing information may have meaning for purposes of judging automaticity, fluency, or motivation. In other tasks or contexts, it may be meaningless. In what types of tasks and contexts response latency might produce meaningful information needs research, including whether such information is more meaningful for formative than summative contexts?
3. Further information may be collected by applying specific additional instruments. Eye tracking is already routinely used in several psychological experiments, and could be applied in TBA for a number of purposes as well. How, and to what extent can one use on screen gaze tracking methods to help computer based training? A lot of specific themes may be proposed. For example, eye tacking may help item development, as problematic elements of the presentation of an item can be identified, in this way. Certain cognitive processes students apply when solving problems can also be identified. Validity issues may be examined in this way as well.
4. Devising general methods for the analysis of person-material interaction. Developing methods of analyzing “trace data” or “interaction data” is important. Many research proposals comment that it is possible to capture a great deal of information about student interactions with material but there are few examples of systematic approaches to data consolidation and analysis. There are approaches used in communication engineering, they are worth to study from the perspective of TBA as well. How the ways of traditionally analyses of social science data can be extended by using these innovative data collection technologies. Such simplified descriptive information (called fingerprints) from trace information (in this case the detailed codes of video records of classrooms) were collected in the TIMSS Video study. The next step is to determine what characteristics of trace data are worth to look at because they are indications of the quality of student learning.

5. A further research theme is the correspondence between assessment frameworks and the actual items presented in the process of computerized testing. Based on the information identified in points 1-4, new methods can be devised to check this correspondence.
6. A general theme for further research is the comparability of results of traditional paper-based testing and results of technology-based assessment. This question may be especially relevant when comparison is one of the main aspects of the assessment, e.g. when trends are established, or in longitudinal research when personal developmental trajectories are studied. What kind of data-collection strategies help linking in such cases.
7. A more general issue is the transfer of knowledge and skill measured by technology. How far skills demonstrated in specific technology-rich environments transfer to other areas, contexts and situations, where not the same technology is present? How skills assessed in simulated environment transfer real-life situations? (See Baker, Niemi, & Chung, 2008, for further discussion.)
8. Whether there are conditions under which formative information can be used for summative purposes without corrupting the value of the formative assessments. Students and teachers should know when they are being judged for consequential purposes. If selected classroom learning sessions are designated as "live" for purposes of collecting summative information, does that reduce the effectiveness of the learning session or otherwise affect the behavior of the student or teacher in important ways?
9. Assessing group outcomes as opposed to individual outcomes. Outcomes of collaboration does not only depend on the communication skills and social/personal skills of the persons involved, as Scardamalia and Bereiter have pointed out in the context of knowledge building as a focus of collaboration. Often, in real life, a team of knowledge workers working on the same project do not come from the expertise background, do not possess the same set of skills and they contribute in different ways to achieve the final outcome. Individuals also gain important learning through the process, but they probably learn different things as well, though there are of course overlaps. How group outcomes could be measured, and what kinds of group outcomes would be important to measure?
10. How, and whether, to account for the contributions of the individual to collaborative activities. Collaboration is an important individual skill but an effective collaboration is in, some sense, best judged by the group's end result. In what types of collaborative technology based tasks might we also be able to gather evidence of the contributions of individuals and what might that evidence be?
11. How is the development of individual outcomes related to group outcomes, and how does this interact with learning task design? Traditionally in education, the learning outcomes expected of everyone at the basic education level is the same – that forms the curriculum standards. Does group productivity require a basic set of core competence from everyone in the team? Answers to these two questions would have important implications for learning design in collaborative settings.
12. Are there interactions with demographic groups for measures such as latency, individual collaborative skills, the collection of summative information from formative learning sessions, or participation in complex assessments such that the meaning of the measures is different for one vs. another group? More precisely, such measures as latency, individual collaborative skills, summative information from formative sessions, etc. have the same meaning in different demographic groups. For example, latency may have a different meaning for males vs. females of a particular country or culture because one group habitually is more careful than the other.

13. How environments in which collaborative skills are measured can be standardized? Can one or all partners in a collaborative situation be replaced by “virtual” partners? Can collaborative activities, contexts, partners be simulated? Can collaborative skills measured in a virtual group, where tested individuals face standardized collaboration-like challenges?
14. Social network analysis; investigating the way people interact with each other when they jointly work on a computer-based task. In network-based collaborative work interactions may be logged, e.g. recording with whom students interact when seeking help, and how these interactions are related to learning. Network analysis software may be used to investigate the interactions among people working on computer-based tasks, and this could provide insights into collaboration. The methods of social network analysis have developed significantly in recent years and can be used to process large numbers of interactions.
15. Automated scoring. On the one hand, recently a lot of research has been carried out on automated scoring (see Williamson, Mislevy, & Bejar, 2006). On the other hand, in practice, real-time automated scoring is used mostly in specific testing situations or is restricted to certain simple item types. Further empirical research is needed e.g. to devise multiple scoring systems, to determine which scoring methods are more broadly applicable, how different scoring methods work in different testing context.
16. One of the possibilities offered by computer-based assessment is for students to be able to save information products for scoring/rating/grading on multiple criteria. An area for research is to investigate how raters grade such complex information products. There is some understanding of how raters grade constructed responses in paper-based assessments and information products can be regarded as complex constructed responses. A related development issue is whether it might be possible to score/rate information products using computer technology. Computer-based assessment has made it possible to store and organize information products for grading but most of the time human raters are required. Tasks involved in producing information products scale differently from single task items. A related but further issue is investigating the dimensionality of computer-based assessment tasks.
17. How could the information gathered by the innovative technology-supported methods (see points 1-3) be utilized to develop new types of adaptive testing in low-stake, formative or diagnostic context? E.g. can additional contextual information be utilized to guide item selection processes?
18. Affective aspects of CBA. It is often assumed that people uniformly enjoy learning in rich technology environments but there is evidence that some people prefer learning using static stimulus material. The research issue would not just be about person-environment fit but would examine how interest changes as people work through tasks in different assessment environments.
19. Measuring emotions. How, and to what extent webcam based emotion detection can be applied? How information gathered by such instruments can be utilized in item development? How measurement of emotions can be utilized in relation to measurement of other domains or constructs, e.g. collaborative skills, social skills.
20. Measuring affective outcomes. Should more general affective outcomes such as ethical behavior in cyberspace be included in the assessment? If so, how can this be done?
21. How computer games can be used for assessment, especially for formative assessment. What is the role of the assessment in games? Where is overlapping between edutainment



and assessment? How can technologies applied in computer games be transferred to assessment? How to detect an addiction to games? How to prevent game addictions?

22. How methods and research results of cognitive/educational neuroscience can be utilized in computer based assessments? For example, how, and to what extent can brainwave detector be used in measuring tiredness and level of concentration?
23. Longitudinal assessment data to build up model(s) of developmental trajectories in 21st century skills. What kind of design will facilitate the building of models of learners' developmental trajectories in the new learning outcome domains? How can technology support collecting, storing and analyzing longitudinal data?
24. Assessment tools for self-assessment v.s. external assessment. Assessment tools should also be an important resource to support learning. When the assessment is conducted by external agencies, especially in the case of high stake assessment, whether these are made on the basis of analysis of interaction data or information products (in which case the assessment is often done through the use of rubrics), the assessment is supported by a team of assessment experts. However, how can such tools be made accessible to teachers (and even students) for learning support through timely and appropriate feedback is important?

## References

- ACT. COMPASS. <http://www.act.org/compass/>
- Ainley, J., Eveleigh, F., Freeman, C., & O'Malley, K. (2009). *ICT in the Teaching of Science and Mathematics in Year 8 in Australia: A Report from the SITES Survey*. Canberra: Department of Education, Employment and Workplace Relations.
- Ainley, M. (2006). Connecting with learning: motivation, affect and cognition in interest processes. *Educational Psychology Review*, 18 (4), 391-405
- American Psychological Association (APA). (1986). *Guidelines for Computer-based Tests and Interpretations*. Washington, D.C.: Author.
- Anderson, R. & Ainley, J. (2009, in press). Technology and learning: Access in schools around the world. In B. McGaw, E. Baker, and P. Peterson *International Encyclopedia of Education*, 3rd Edition. Amsterdam: Elsevier.
- Baker E. L., Niemi, D., & Chung, G. K. W. K. (2008). Simulations and the transfer of problem-solving knowledge and skills. In E. Baker, J. Dickerson, W. Wulfeck, & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations*, (pp. 1-17.). New York: Lawrence Erlbaum Associates.
- Ball, S. et al. (2006): Accessibility in e-Assessment Guidelines Final Report. Commissioned by TechDis for the E-Assessment Group and Accessible E-Assessment. Report Prepared by Edexcel. Available: [http://www.techdis.ac.uk/resources/files/Final%20report%20\(TechDis\)SBfinal.pdf](http://www.techdis.ac.uk/resources/files/Final%20report%20(TechDis)SBfinal.pdf)
- Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning and Assessment*, 2(3). Available: <http://www.bc.edu/research/intasc/jtla/journal/v2n3.shtml>
- Bennett, R. E. (2001). How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archives*, 9(5). Available: <http://epaa.asu.edu/epaa/v9n5.html>
- Bennett, R. E. (2006). Moving the field forward: Some thoughts on validity and automated scoring. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 403-412). Mahwah, NJ: Erlbaum.

## Assessment and Teaching of 21st Century Skills project white papers

- Bennett, R. E. (2009). *A critical look at the meaning and basis of formative assessment* (RM-09-06). Princeton, NJ: Educational Testing Service.
- Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, 17(4), 9-17.
- Bennett, R. E., Braswell, J., Oranje, A., Sandene, B, Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment*, 6(9). Available: <http://escholarship.bc.edu/jtla/vol6/9/>
- Bennett, R. E., Braswell, J., Oranje, A., Sandene, B, Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment*, 6(9). Available: <http://escholarship.bc.edu/jtla/vol6/9/>
- Bennett, R. E., Goodman, M., Hessinger, J., Liggett, J., Marshall, G., Kahn, H., & Zack, J. (1999). Using multimedia in large-scale computer-based testing programs. *Computers in Human Behavior*, 15, 283-294.
- Bennett, R. E., Morley, M., & Quardt, D. (2000). Three response types for broadening the conception of mathematical problem solving in computerized tests. *Applied Psychological Measurement*, 24, 294-309.
- Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project (NCES 2007-466). Washington, DC: National Center for Education Statistics, US Department of Education. Available: <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007466>
- Bennett, R.E., Jenkins, F., Persky, H., & Weiss, A. (2003). Assessing complex problem-solving performances. *Assessment in Education*, 10, 347-359.
- Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *Journal of Technology, Learning and Assessment*, 6(1). Available: <http://escholarship.bc.edu/jtla/vol6/1/>
- Bergholtz, M., Grégoire, B., Johannesson, P., Schmitt, M., Wohed, P. & Zdravkovic, J. (2005). Integrated Methodology for linking business and process models with risk mitigation. International Workshop on Requirements Engineering for Business Need and IT Alignment (REBNITA 2005), Paris, August 2005. [http://efficient.citi.tudor.lu/cms/efficient/content.nsf/0/4A938852840437F2C12573950056F7A9/\\$file/Rebnita05.pdf](http://efficient.citi.tudor.lu/cms/efficient/content.nsf/0/4A938852840437F2C12573950056F7A9/$file/Rebnita05.pdf)
- Berglund, A., Boag, S., Chamberlin, D., Fernández, M., Kay, M., Robie, J. & Siméon, J. (Eds) (2007). XML Path Language (XPath) 2.0. W3C Recommendation 23 January 2007. <http://www.w3.org/TR/2007/REC-xpath20-20070123/>
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web: A new form of web that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284, 34-43
- Bernstein, J. (1999). *PhonePass testing: Structure and construct*. Menlo Park, CA: Ordinate.
- Bernstein, H. (2000). Recent changes to RasMol, recombining the variants. *Trends in Biochemical Sciences (TIBS)*, 25 (9) 453-455.
- Blech, C., & Funke, J. (2005). Dynamis review: An overview about applications of the Dynamis approach in cognitive psychology. Bonn: Deutsches Institut für Erwachsenenbildung. Available: [http://www.die-bonn.de/esprid/dokumente/doc-2005/blech05\\_01.pdf](http://www.die-bonn.de/esprid/dokumente/doc-2005/blech05_01.pdf)
- Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Tyler (Ed.), *Educational evaluation: New roles, new means*. The 63rd yearbook of the National Society for the Study of Education, part 2 (Vol. 69) (p. 26-50). Chicago: University of Chicago Press.

- Booth, D. & Liu, K. (Eds.) (2007). Web Services Description Language (WSDL) Version 2.0 Part 0: Primer. W3C Recommendation 26 June 2007. <http://www.w3.org/TR/2007/REC-wsdl20-primer-20070626>
- Boud, D., Cohen, R., & Sampson, J. (1999). Peer learning and assessment. *Assessment & Evaluation in Higher Education*, 24(4), 413-426.
- Bray, T., Paoli, J., Sperberg-McQueen, C., Maler, E. & Yergeau, F. (2008). (Eds.), *Extensible Markup Language (XML) 1.0* (Fifth Edition) W3C Recommendation 26 November 2008. <http://www.w3.org/TR/2008/REC-xml-20081126/>
- Bray, T., Paoli, J., Sperberg-McQueen, C., Maler, E. & Yergeau, F., Cowan, J., (2006). (Eds.). XML 1.1 (Second Edition), W3C Recommendation, 16 August 2006. <http://www.w3.org/TR/2006/REC-xml11-20060816/>
- Brickley, D., & Guha, R. (2004). RDF vocabulary description language 1.0: RDF Schema. *W3C Recommendation*. <http://www.w3.org/TR/2004/REC-rdf-schema-20040210/>
- Bridgeman, B. (2009). Experiences from Large-Scale Computer-Based Testing in the USA. In F. Scheuermann, & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 39-44). Luxembourg: Office for Official Publications of the European Communities.
- Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education*, 16, 191-205.
- Carlisle, D., Ion, P., Miner, R. & Poppelier, N. (Eds.) (2003). Mathematical Markup Language (MathML) Version 2.0 (Second Edition). W3C Recommendation 21 October 2003. <http://www.w3.org/TR/2003/REC-MathML2-20031021/>
- Chatty, S., Sire, S., Vinot J.-L., Lecoanet, P., Lemort, A. & Mertz, C. (2004). *Revisiting visual interface programming: creating GUI tools for designers and programmers*. Proc UIST'04, October 24-27, 2004 Santa Fe, NM, USA. ACM Digital Library.
- Clement, L., Hatley, A., von Riegen, C. & Rogers, T. (2004) *UDDI Version 3.0.2, UDDI Spec Technical Committee Draft, Dated 20041019*. Organization for the Advancement of Structured Information Standards (OASIS). <http://uddi.org/pubs/uddi-v3.0.2-20041019.htm>
- Clyman, S. G., Melnick, D. E., & Clauser, B. E. (1995). Computer-based case simulations. In L. Mancall & P. G. Bashook (Ed.), *Assessing clinical reasoning: The oral examination and alternative methods* (pp. 139-149). Evanston, IL: American Board of Medical Specialties.
- Carnegie Learning. *Cognitive Tutors*. <http://www.carnegielearning.com/products.cfm>
- Catts, R. & Lau, J. (2008). *Towards Information Literacy Indicators*. Paris:UNESCO.
- Conole, G., & Waburton, B. (2005). A review of computer-assisted assessment. *ALT-J, Research in Learning Technology*, 13(1), 17-31
- College Board. *ACCUPLACER*. <http://www.collegeboard.com/student/testing/accuplacer/>
- Corbiere, A., (2008). A Framework to Abstract The Design Practices of e-Learning System Projects in IFIP International Federation for Information Processing, Volume 275; Open Source Development, Communities and Quality; Barbara Russo, Ernesto Damiani, Scott Hissam, Björn Lundell, Giancarlo Succi; (pp. 317–323). Boston: Springer.
- Cost, R., Finin, T., Joshi, A., Peng, Y., Nicholas, C., Soboroff, I., Chen, H., Kagal, L., Perich, F., Zou, Y., Tolia, S. (2002). ITalks: A Case Study in the Semantic Web and DAML+OIL, *IEEE Intelligent Systems*, 17, (1) 40-47.
- Cross, R. (2004a). Review of item banks. In Sclater, N (Ed.), *Final report for the Item Bank Infrastructure Study (IBIS)*, (pp. 17-34). Bristol: JISC.

- Cross, R. (2004b). Metadata and searching. In Sclater, N (Ed.), *Final report for the Item Bank Infrastructure Study (IBIS)*, (pp. 87-102). Bristol: JISC.
- Csapó B., Molnár G., & R. Tóth, K. (2009). Comparing paper-and-pencil and online assessment of reasoning skills. A pilot study for introducing electronic testing in large-scale assessment in Hungary. In F. Scheuermann, & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 113-118). Luxemburg: Office for Official Publications of the European Communities.
- CTB/McGraw-Hill. *Acuity*.  
[http://www.ctb.com/products/product\\_summary.jsp?FOLDER%3C%3Efolder\\_id=1408474395292638](http://www.ctb.com/products/product_summary.jsp?FOLDER%3C%3Efolder_id=1408474395292638)
- Decker, S., Melnik, S., Van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., Erdmann, M., Horrocks, I. (2000). The Semantic Web: The Roles of XML and RDF, *IEEE Internet Computing*, **15**, (5) 2-13
- Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. *Learning in Humans and Machine: Towards an interdisciplinary learning science*, 189-211.
- Draheim, D., Lutteroth, C. & Weber G. (2006). Graphical user interface as documents. In CHINZ 2006 – Design Centred HCI, July 6-7, 2006, Christchurch, New Zealand. ACM digital library.
- Dragow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 471-515). Westport, CT: American Council on Education/Praeger.
- EMB (Education and Manpower Bureau HKSAR) (2001). Learning to Learn - The Way Forward in Curriculum. Retrieved 11/09/2009. from <http://www.edb.gov.hk/index.aspx?langno=1&nodeID=2877>.
- Educational Testing Service (ETS). *Graduate Record Examinations (GRE)*.  
<http://www.ets.org/portal/site/ets/menuitem.fab2360b1645a1de9b3a0779f1751509/?vgnnextoid=b195e3b5f64f4010VgnVCM10000022f95190RCRD>
- Educational Testing Service (ETS). *Test of English as a Foreign Language iBT (TOEFL iBT)*.  
[http://www.ets.org/portal/site/ets/menuitem.fab2360b1645a1de9b3a0779f1751509/?vgnnextoid=69c0197a484f4010VgnVCM10000022f95190RCRD&WT.ac=Redirect\\_ets.org\\_toefl](http://www.ets.org/portal/site/ets/menuitem.fab2360b1645a1de9b3a0779f1751509/?vgnnextoid=69c0197a484f4010VgnVCM10000022f95190RCRD&WT.ac=Redirect_ets.org_toefl)
- Educational Testing Service (ETS). *TOEFL Practice Online*. <http://toeflpractice.ets.org/>
- Eggen, T. & Straetmans, G. (2009). Computerised adaptive testing at the entrance of primary school teacher training college. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing*, (pp. 134-144). Luxemburg: Office for Official Publications of the European Communities.
- Farcot, M. & Latour, T. (2009). Transitioning to Computer-Based Assessments: A Question of Costs. In F. Scheuermann, & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*, (pp. 108-116). Luxemburg: Office for Official Publications of the European Communities.
- Ferraiolo, J., Jun, J. & Jackson, D. (2009). Scalable Vector Graphics (SVG) 1.1 Specification. W3C Recommendation 14 January 2003, edited in place 30 April 2009.  
<http://www.w3.org/TR/2003/REC-SVG11-20030114/>
- Feurzeig, W., & Roberts, N. (1999). *Modeling and simulation in science and mathematics education*: Springer Verlag.
- Flores, F., Quint, V. & Vatton, I. (2006). Templates, Microformats and Structured Editing. *Proceedings of DocEng'06, ACM Symposium on Document Engineering, 10-13 October 2006*, (pp. 188-197) Amsterdam, The Netherlands.

- Gašević, D., Jovanović, J. & Devedžić, V. (2004). Ontologies for creating learning object content. In M. Gh. Negoita et al. (Eds.), *KES 2004, LNAI 3213*, pp. 284–291.
- Graduate Management Admission Council (GMAC). *Graduate Management Admission Test (GMAT)*. <http://www.mba.com/mba/thegmat>
- Greiff, S., & Funke, J. (2008). Measuring complex problem solving: The MicroDYN approach. Heidelberg: unpublished manuscript. Available: [http://www.psychologie.uni-heidelberg.de/ae/allg/forschun/dfg\\_komp/Greiff&Funke\\_2008\\_MicroDYN.pdf](http://www.psychologie.uni-heidelberg.de/ae/allg/forschun/dfg_komp/Greiff&Funke_2008_MicroDYN.pdf)
- Gruber, T. (1991 April). The Role of Common Ontology in Achieving Sharable, Reusable Knowledge Bases, *Proc. Second Int'l Conf. Principles of Knowledge Representation and Reasoning*, (pp. 601-602). Cambridge, MA: Morgan Kaufmann Publishers.
- Grubber, T. (1993). A Translation Approach to Portable Ontology Specifications, *Knowledge Acquisition*, 5, 199-220.
- Guarino, N. & Giaretta P. (1995). Ontologies and knowledge bases: Towards a Terminological Clarification, In N. Mars (Ed.), *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, (pp. 25-32). Amsterdam: IOS Press.
- Gudgin, M., Hadley, M., Mendelsohn, N., Moreau, J.-J., Nielsen, H., Karmarkar, A. & Lafon, Y. (Eds.) (2007). *SOAP Version 1.2 Part 1: Messaging Framework (Second Edition)*. W3C Recommendation 27 April 2007. <http://www.w3.org/TR/2007/REC-soap12-part1-20070427/>
- Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of Educational Computing Research*, 17(4), 397 -431.
- Hadwin, A., Winne, P., & Nesbit, J. (2005). Roles for software technologies in advancing research and theory in educational psychology. *British Journal of Educational Psychology*, 75, 1-24.
- Haldane, S. (2009). Delivery platforms for national and international computer based surveys. In F. Sheuermann & J. Björnsson (Eds). *The transition to Computer-Based Assessment: New approaches to skills assessment and implications for large-scale testing*, (pp. 63-67). Luxembourg: Office for Official Publications of the European Communities.
- Halldórsson, A., McKelvie, P., & Björnsson, J. (2009). Are Icelandic boys really better on computerized tests than conventional ones: Interaction between gender test modality and test performance. In F. Sheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing*, (pp. 178-193). Luxembourg: Office for Official Publications of the European Communities.
- Hendler, J. (2001). Agents and the Semantic Web, *IEEE Intelligent Systems*, 16, (2) 30-37.
- Henri, F. (1992). Computer conferencing and content analysis. In A. R. Kaye (Ed.), *Collaborative Learning Through Computer Conferencing* (pp. 117-136). Berlin: Springer-Verlag.
- Herráez, A. (2007). *How to use Jmol to study and present molecular structures*, Vol.1. Morrisville, NC: Lulu Enterprises.
- Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment*, 5(2). Available: <http://escholarship.bc.edu/jtla/vol5/2/>
- IEEE LTSC (2002). IEEE Standard for Learning Object Metadata. Computer Society/Learning Technology Standards Committee.
- IMS (2006). IMS Question and Test Interoperability Overview, Version 2.0 Final Specification. IMS Global Learning Consortium, Inc. Available: [http://www.imsglobal.org/question/qti\\_v2p0/imsqti\\_oviewv2p0.html](http://www.imsglobal.org/question/qti_v2p0/imsqti_oviewv2p0.html)
- International ICT Literacy Panel (Educational Testing Service) (2002). *Digital Transformation: A Framework for ICT Literacy*. Princeton, NJ: Educational Testing Service.

- ISO/IEC-10746-1 (1998). Open Distributed Processing Reference Model, Part 1: Overview. ISO/IEC JTC1 SC7.
- Jadoul, R. & Mizohata, S. (2006). PRECODEM, an example of TAO in service of employment. IADIS International Conference on Cognition and Exploratory Learning in Digital Age, CELDA 2006, 8-10 December 2006. Barcelona, Spain.  
[https://www.tao.lu/downloads/publications/CELD2006\\_PRECODEM\\_paper.pdf](https://www.tao.lu/downloads/publications/CELD2006_PRECODEM_paper.pdf)
- Jadoul, R. & Mizohata, S. (2007). Development of a Platform Dedicated to Collaboration in the Social Sciences. Oral presentation at IADIS International Conference on Cognition and Exploratory Learning in Digital Age, CELDA 2007, 7-9 December 2007. Carvoeiro, Portugal.  
[https://www.tao.lu/downloads/publications/CELD2007\\_Development\\_of\\_a\\_Platform\\_paper.pdf](https://www.tao.lu/downloads/publications/CELD2007_Development_of_a_Platform_paper.pdf)
- Jadoul, R., Plichart, P., Swietlik, J. & Latour, T. (2006). eXULiS - a Rich Internet Application (RIA) framework used for eLearning and eTesting. IV International Conference On Multimedia And Information And Communication Technologies In Education, m-ICTE 2006. 22-25 November, 2006. Seville, Spain In: Méndez-Vilas, A., Solano Martin, A., Mesa González, J., Mesa González, J.A. (eds.): Current Developments in Technology-Assisted Education, Vol. 2. FORMATEX, Badajoz, Spain, (2006) pp. 851-855.  
<http://www.formatex.org/micte2006/book2.htm>
- Johnson, M. & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning, and Assessment*, 4 (5).
- Kamareddine, F., Lamar, R., Maarek, M. & Wells, J. (2007). Restoring natural language as a computerized mathematics input method. In M. Kauers et al. (Eds.), MKM/Calculemus 2007, LNAI 4573, pp. 280-295. [http://dx.doi.org/10.1007/978-3-540-73086-6\\_23](http://dx.doi.org/10.1007/978-3-540-73086-6_23)
- Kamareddine, F., Maarek, M., Retel, K. & Wells, J. (2007) Narrative structure of mathematical texts. In M. Kauers et al. (Eds.), MKM/Calculemus 2007, LNAI 4573, pp. 296-312.  
[http://dx.doi.org/10.1007/978-3-540-73086-6\\_24](http://dx.doi.org/10.1007/978-3-540-73086-6_24)
- Kay, M. (Ed.) (2007). XSL Transformations (XSLT) Version 2.0. W3C Recommendation 23 January 2007. <http://www.w3.org/TR/2007/REC-xslt20-20070123/>.
- Kelly, M. & Haber, J. (2006). *National Educational Technology Standards for Students (NETS\*S): Resources for Assessment*. Eugene, OR: The International Society for Technology and Education.
- Kerski, J. (2003). The implementation and effectiveness of geographic information systems technology and methods in secondary education. *Journal of Geography*, 102(3), 128-137.
- Khang, J. & McLeod, D. (1998). Dynamic Classificational Ontologies: Mediation of information sharing in cooperative federated database systems. In M. P. Papazoglou, G. Sohlager (Eds.), *Cooperative Information Systems: Trends and direction*, (pp. 179-203). San Diego, CA: Academic Press.
- Kia, E., Quint, V. & Vatton, I. (2008). XTiger Language Specification. Available: <http://www.w3.org/Amaya/Templates/XTiger-spec.html>.
- Klyne, G., & Carrol, J. (2004). Resource description framework (RDF): Concepts and abstract syntax. *W3C Recommendation*. <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/>.
- Koretz, D. (2008). *Measuring up. What educational testing really tells us*. Cambridge, MA: Harvard University Press.
- Kyllonen, P. & Lee, S. (2005). Assessing problem solving in context. In O. Wilhelm and R. Engle (Eds.) *Handbook of Understanding and Measuring Intelligence*. (pp 11-25) Thousand Oaks, CA: Sage.

- Kyllonen, P. (2009). New constructs, methods and directions for computer-based assessment. In F. Sheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 151-156). Luxembourg: Office for Official Publications of the European Communities.
- Latour, T. & Farcot, M. (2008). An Open Source and Large-Scale Computer-Based Assessment Platform: A real Winner. In F. Scheuermann & A. Guimaraes Pereira (Eds.), *Towards a research agenda on Computer-Based Assessment. Challenges and needs for European educational measurement*, (pp. 64-67). Luxembourg: Office for Official Publications of the European Communities.
- Laubscher, R., Olivier, M. S., Venter, H. S., Eloff, J. H., and Rabe, D. J. (2005). The role of key loggers in computer-based assessment forensics. In *Proceedings of the 2005 Annual Research Conference of the South African institute of Computer Scientists and information Technologists on IT Research in Developing Countries* (White River, South Africa, September 20 - 22, 2005). SAICSIT, vol. 150. South African Institute for Computer Scientists and Information Technologists, 123-130.
- Lave J. (1988). *Cognition in Practice*. Cambridge: Cambridge University Press
- Law, N. (2005). Assessing Learning Outcomes in CSCL Settings. In T.-W. Chan, T. Koschmann & D. Suthers (Eds.), *Proceedings of the Computer Supported Collaborative Learning Conference (CSCL) 2005* (pp. 373-377). Taipei: Lawrence Erlbaum Associates.
- Law, N., Yuen, H. K., Shum, M., & Lee, Y. (2007). *Phase (II) Study on Evaluating the Effectiveness of the 'Empowering Learning and Teaching with Information Technology' Strategy (2004/2007) Final Report*. Hong Kong: Hong Kong Education Bureau.
- Lehtinen, E., Hakkarainen, K., Lipponen, L., Rahikainen, M., & Muukkonen, H. (1999). *Computer supported collaborative learning: A review. Computer supported collaborative learning in primary and secondary education*. A final report for the European Commission, Project, 1-46.
- Lennon, M., Irwin K., Von Davier, M., Wagner, M. & Yamamoto, K. (2003). *Feasibility Study for the PISA ICT Literacy Assessment, Report to Network A*. Paris: OECD.
- Lie, H. & Bos, B. (2008). Cascading Style Sheets, level 1. W3C Recommendation 17 Dec 1996, revised 11 Apr 2008. <http://www.w3.org/TR/2008/REC-CSS1-20080411>
- Linn, M., & Hsi, Sherry (1999). *Computers, teachers, peers : science learning partners*. Mahwah, N.J.: Lawrence Erlbaum Associates.
- Longley, P. (2005). *Geographic information systems and science*: Wiley.
- Lőrincz, A. (2008). Machine situation assessment and assistance: Prototype for severely handicapped children. In A. K. Varga, J. Vásárhelyi, & L. Samuelis, (Eds).. In *Proceedings of Regional Conference on Embedded and Ambient Systems, Selected Papers* (pp 61-68), Budapest: John von Neumann Computer Society. Available: [http://nippg.inf.elte.hu/index.php?option=com\\_remository&Itemid=27&func=fileinfo&id=155](http://nippg.inf.elte.hu/index.php?option=com_remository&Itemid=27&func=fileinfo&id=155)
- Macdonald, J. (2003). Assessing online collaborative learning: process and product. *Computers & Education*, 40(4), 377-391.
- Maedche, A. & Staab, S. (2001). Ontology Learning for the Semantic Web, *IEEE Intelligent Systems*, 16, (2) 72-79
- Mahalingam, K., & Huns, M. (1997). An Ontology Tool for Query Formulation in An Agent-Based Context, *Proc. Second IFCIS Int'l Conf. Cooperative Information Systems*, Kiawah Island, South Carolina, USA; June, IEEE Computer Society, pp. 170-178
- Marks, A., & Cronje, J. (2008). Randomised Items in Computer-based Tests: Russian Roulette in Assessment? *Journal of Educational Technology & Society*, 11(4), 41-50.

## Assessment and Teaching of 21st Century Skills project white papers

- Markauskaite, L. (2007). Exploring the structure of trainee teachers' ICT literacy: the main components of, and relationships between, general cognitive and technical capabilities. *Education Technology Research Development* 55: 547-572.
- Martin, M., Mullis, I., & Foy, P. (2008). *TIMSS 2007 International Science Report. Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eight Grades*. Chestnut Hill, MA: IEA TIMSS & PIRLS International Study Center.
- Martin, R., Busana, G. & Latour, T. (2009). Vers une architecture de testing assisté par ordinateur pour l'évaluation des acquis scolaires dans les systèmes éducatifs orientés sur les résultats. In Jean-Guy Blais (Ed.), *Évaluation des apprentissages et technologies de l'information et de la communication, Enjeux, applications et modèles de mesure*, (pp. 13-34). Quebec: Presses de l'Université Laval.
- McConnell, D. (2002). The experience of collaborative assessment in e-learning. *Studies in continuing education*, 24(1), 73-92.
- McDaniel, M., Hartman, N., Whetzel, D., & Grubb, W. (2007). Situational judgment tests: response, instructions and validity: A meta-analysis. *Personnel Psychology*, 60, 63-91.
- McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers and Education*, 39 (3), 299-312.
- Means, B., & Haertel, G. (2002). Technology supports for assessing science inquiry. In N. R. Council (Ed.), *Technology and Assessment: Thinking Ahead: Proceedings from a Workshop* (pp. 12-25). Washington, DC: National Academy Press.
- Means, B., Penuel, B., & Quellmalz, E. (2000). Developing assessments for tomorrow's classrooms. Paper presented at the The Secretary's Conference on Educational Technology 2000. Retrieved 19/09/2009, from <http://tepservers.ucsd.edu/courses/tep203/fa05/b/articles/means.pdf>
- Mead, A., D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114, 449-458.
- Mellar, H., Bliss, J., Boohan, R., Ogborn, J., & Tompsett, C. (Ed.). (1994). *Learning with Artificial Worlds: Computer Based Modelling in the Curriculum*. London: The Falmer Press.
- Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.
- Microsoft. Extensible Application Markup Language (XAML). <http://msdn.microsoft.com/en-us/library/ms747122.aspx>
- Miller, J. & Mukerji, J. (Eds.) (2003) MDA Guide Version 1.0.1. Object Management Group. <http://www.omg.org/cgi-bin/doc?omg/03-06-01.pdf>
- Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA) (1999). *National Goals for Schooling in the Twenty First Century*. Curriculum Corporation: Melbourne.
- Ministerial Council on Education, Early Childhood Development and Youth Affairs (MCEECDYA) (2008). *Melbourne Declaration on Education Goals for Young Australians*. Curriculum Corporation: Melbourne.
- Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA) (2000) *Learning in an Online World: the School Education Action Plan for the Information Economy*. Adelaide: Education Network Australia.
- Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA) (2005) *Contemporary Learning: Learning in an On-line World*. Carlton, Vic: Curriculum Corporation.
- Ministerial Council for Education, Employment, Training and Youth Affairs (MCEETYA) (2007). *National Assessment Program - ICT Literacy Years 6 & 10 Report*, Carlton, Vic.: Curriculum Corporation.



- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). A brief introduction to Evidence-Centered Design. (CSE Report 632). Los Angeles, CA: UCLA CRESST.
- Mislevy, R. J., Almond, R. G., Steinberg, L. S., & Lukas, J. F. (2006). Concepts, terminology, and basic models in evidence-centered design. In D. M., Williamson, R. J., Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15-47). Mahwah, NJ: Erlbaum.
- Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25(4), 6-20.
- Mozilla Foundation. XML User Interface Language. [https://developer.mozilla.org/en/XUL\\_Reference](https://developer.mozilla.org/en/XUL_Reference)
- Mullis, I., Martin, M., & Foy, P. (2008). *TIMSS 2007 International Mathematics Report. Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eight Grades*. Chestnut Hill, MA: IEA TIMSS & PIRLS International Study Center.
- Mullis, I., Martin, M., Kennedy, A., & Foy P. (2007). *PIRLS 2006 International Report: IEA's Progress in International Reading Literacy Study in Primary School on 40 countries*. Chestnut Hill, MA: Boston College.
- Northwest Evaluation Association. *Measures of Academic Progress MAP*. <http://www.nwea.org/products-services/computer-based-adaptive-assessments/map>
- OECD (2008). Issues arising from the PISA 2009 field trial of the assessment of reading of electronic texts. Document of the 26<sup>th</sup> meeting of the PISA Governing Board. Paris, OECD Directorate for Education.
- OECD (2009). PISA CBAS analysis and results – Science performance on paper and pencil and electronic tests. Paris: OECD.
- OMG. The object Management Group. <http://www.omg.org/>
- Oregon Department of Education. *Oregon Assessment of Knowledge and Skills (OAKS)*. <http://www.oaks.k12.or.us/resourcesGeneral.html>
- Organisation for Economic Co-operation and Development (OECD) (2007). *PISA 2006 Science Competencies for Tomorrow's World*. Paris: OECD.
- Organisation for Economic Co-operation and Development (OECD) (2008). *The OECD Programme for the Assessment of Adult Competencies (PIAAC)*. Paris: OECD.
- Patel-Schneider P., Hayes P., & Horrocks, I. (2004). OWL Web Ontology Language Semantics and Abstract Syntax. *W3C Recommendation*. <http://www.w3.org/TR/2004/REC-owl-semantic-20040210/>
- Pea, R. (2002). *Learning science through collaborative visualization over the Internet*. Paper presented at the Nobel Symposium (NS 120), Stockholm, Sweden.
- Pearson. *PASeries*. <http://education.pearsonassessments.com/pai/ea/products/paseries/paseries.htm>
- Pellegrino, J., Chudowosky, N. & Glaser, R. (2004). *Knowing What Students Know: the Science and Design of Educational Assessment*. Washington, DC: National Academy Press.
- Pelgrum, W. (2008). School Practices and Conditions for Pedagogy and ICT. In N. Law, W. Pelgrum & T. Plomp (Eds.), *Pedagogy and ICT use in schools around the world: Findings from the IEA SITES 2006 study*. Hong Kong: CERC and Springer.
- Plichart P., Jadoul R., Vandenabeele L., & Latour T. (2004). TAO, a Collective distributed computer-based assessment framework built on semantic web standards. In Proceedings of the International Conference on Advances in Intelligent Systems – Theory and Application AISTA2004, In cooperation with IEEE Computer Society, November 15-18, 2004. Luxembourg, Luxembourg.

## Assessment and Teaching of 21st Century Skills project white papers

- Plichart, P., Latour, T., Busana, G., & Martin, R. (2008). Computer based school system monitoring with feedback to teachers. In Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2008 (pp. 5065-5070). Chesapeake, VA: AACE.
- Plomp, T., Anderson, R. E., Law, N., & Quale, A. (Eds.). (2009). *Cross-national Information and Communication Technology Policy and Practices in Education* (2nd ed.). Greenwich, CT: Information Age Publishing Inc.
- Poggio, J., Glasnapp, D., Yang, X., & Poggio, A. (2004). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program (2004) *Journal of Technology, Learning, and Assessment*, 3 (6), 30-38.
- Poole, J. (2001). Model-Driven Architecture: Vision, standards and Emerging technologies. Position paper in Workshop on Metamodeling and Adaptive Object Models, ECOOP 2001, Budapest, Hungary. Available: [http://www.omg.org/mda/mda\\_files/Model-Driven\\_Architecture.pdf](http://www.omg.org/mda/mda_files/Model-Driven_Architecture.pdf)
- Popper, K. (1972). *Objective knowledge: An evolutionary approach*: Oxford University Press, USA.
- Raggett, D., Le Hors, A. & Jacobs, I. (1999). HTML 4.01 Specification. W3C Recommendation 24 December 1999. <http://www.w3.org/TR/1999/REC-html401-19991224>
- Ram, S. & Park, J. (2004). Semantic Conflict Resolution Ontology (SCROL): An Ontology for Detecting and Resolving Data and Schema-Level Semantic Conflicts, *IEEE Trans. Knowledge and Data Eng.*, 16, (2) 189-202
- Quellmalz, E., & Haertel, G. (2004). Use of technology-supported tools for large-scale science assessment: Implications for assessment practice and policy at the state level: Committee on Test Design for K-12 Science Achievement, Center for Education, National Research Council.
- Quellmalz, E., & Pellegrino, J. (2009). Technology and testing. *Science*, 323(5910), 75.
- Quellmalz, E., Timms, M., & Buckley, B. (2009). Using Science Simulations to Support Powerful Formative Assessments of Complex Science Learning. Paper presented at the American Educational Research Association Annual Conference. Retrieved 11/09/2009, from [http://simscientist.org/downloads/Quellmalz\\_Formative\\_Assessment.pdf](http://simscientist.org/downloads/Quellmalz_Formative_Assessment.pdf)
- Reich, K., & Petter, C. (2009). eInclusion, eAccessibility and design for all issues in the context of European Computer-Based Assessment. In F. Scheuermann, & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 68-73). Luxemburg: Office for Official Publications of the European Communities.
- Sakayauchi, M., Maruyama, H., & Watanabe, R. (2009). National Policies and Practices on ICT in Education: Japan In T. Plomp, R. E. Anderson, N. Law & A. Quale (Eds.), *Cross-national Information and Communication Technology Policy and Practices in Education* (2nd ed.), pp. 441-457. Greenwich, CT: Information Age Publishing Inc.
- Sandene, B., Bennett, R.E., Braswell, J., & Oranje, A. (2005). Online assessment in mathematics. In B. Sandene, N. Horkay, R.E. Bennett, N. Allen, J. Braswell, B. Kaplan, & A. Oranje (Eds.), *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project* (NCES 2005-457). Washington, DC: National Center for Education Statistics, US Department of Education. Retrieved July 29, 2007 from <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457>.
- Sayle, R. & Milner-White, E. (1995). RasMol: Biomolecular graphics for all. *Trends in Biochemical Sciences (TIBS)*, 20 (9) 374.
- Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Smith (Ed.), *Liberal education in a knowledge society*, pp. 67-98. Chicago, IL: Open Court.

- Scardamalia, M., & Bereiter, C. (2003). Knowledge building environments: Extending the limits of the possible in education and knowledge work. In A. DiStefano, K. E. Rudestam & R. Silverman (Eds.), *Encyclopedia of distributed learning*. (pp. 269-272). Thousand Oaks, CA: Sage Publications.
- Scheuermann, F., & Björnsson J. (Eds.) (2009). *New approaches to skills assessment and implications for large-scale testing. The transition to computer-based assessment*. Luxembourg: Office for Official Publications of the European Communities.
- Scheuermann, F., & Guimarães Pereira, A. (Eds.) (2008). *Towards a research agenda on computer-based assessment*. Luxembourg: Office for Official Publications of the European Communities.
- Schulz, W., Fraillon, J., Ainley, J., Losito, B. & Kerr, D. (2008). *International Civic and Citizenship Education Study. Assessment Framework*. Amsterdam: IEA.
- Schmidt, D. C. (2006). Model-Driven Engineering. *IEEE Computer* 39 (2) 25-31
- Schmitt, M. & Grégoire, B., (2006). Business service network design: from business model to an integrated multi-partner business transaction. Joint International Workshop on Business Service Networks and Service oriented Solutions for Cooperative Organizations (BSN-SoS4CO '06), June 2006, San Francisco, California, USA. Available: [http://efficient.citi.tudor.lu/cms/efficient/content.nsf/0/4A938852840437F2C12573950056F7A9/\\$file/Schmitt06\\_BusinessServiceNetworkDesign\\_SOS4CO06.pdf](http://efficient.citi.tudor.lu/cms/efficient/content.nsf/0/4A938852840437F2C12573950056F7A9/$file/Schmitt06_BusinessServiceNetworkDesign_SOS4CO06.pdf)
- Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (p. 39-83). Chicago: Rand McNally.
- Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher*, 27(2), 4.
- Shermis, M. D., & Burstein, J. C. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.
- Singapore Ministry of Education (1997). *Masterplan for IT in Education: 1997-2002* Retrieved 17/8/2009, 2009, from <http://www.moe.gov.sg/edumall/mpite/index.html>
- Singleton, C. (2001). Computer-based assessment in education. *Educational and Child Psychology*, 18 (3), 58-74
- Sowa, J. (2000) Knowledge Representation. Logical, Philosophical, and Computational Foundations, Brooks-Cole, Pacific-Groce, CA, USA.
- Stevens, R. H., & Casillas, A. C. (2006). Artificial neural networks. In D. M., Williamson, R. J., Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 259-311). Mahwah, NJ: Erlbaum.
- Stevens, R. H., Lopo, A. C., & Wang, P. (1996). Artificial neural networks can distinguish novice and expert strategies during complex problem solving. *Journal of the American Medical Informatics Association*, 3, 131-138.
- Suchman, L.A. (1987). *Plans and Situated Actions. The problem of human machine communication*. Cambridge: Cambridge University Press.
- Tan, W., Yang, F., Tang, A., Lin, S. & Zhang, X. (2008). An E-Learning System Engineering Ontology Model on the Semantic Web for Integration and Communication. In F. Li et al. (Eds.). ICWL 2008, LNCS 5145 (pp. 446-456).
- Thompson, N. & Wiess, D. (2009). Computerised and adaptive testing in educational assessment. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 127-133). Luxembourg: Office for Official Publications of the European Communities.

- Tinker, R., & Xie, Q. (2008). Applying Computational Science to Education: The Molecular Workbench Paradigm. *Computing in Science & Engineering*, 10(5), 24-27.
- Tissoires, B. & Conversy, S. (2008). Graphic rendering as a compilation chain. In Graham T. & Palanque, P. (Eds.), DSVIS 2008, LNCS 5136, (pp. 267-280).
- Torney-Purta, J., Lehmann, R., Oswald, H., & Schulz, W. (2001). *Citizenship and Education in Twenty-eight Countries: Civic Knowledge and Engagement at Age Fourteen*. Delft: IEA.
- Turki, S., Aïdonis, Ch., Khadraoui, A. & Léonard, M. (2004). Towards Ontology-Driven Institutional IS Engineering. Open INTEROP Workshop on "Enterprise Modelling and Ontologies for Interoperability", EMOI-INTEROP 2004; Co-located with CaiSE'04 Conference, Riga (Latvia), 7-8 June 2004
- Van der Vet, P. & Mars, N. (1998). Bottom up Construction of Ontologies, *IEEE Trans. Knowledge and Data Eng.*, 10(4) 513-526.
- Vargas-Vera, M. & Lytras, M. (2008). Personalized Learning Using Ontologies and Semantic Web Technologies. In M.D. Lytras et al. (Eds.). WSKS 2008, LNAI 5288, (pp. 177-186).
- Virginia Department of Education. *Standards of Learning Tests*.  
[http://www.doe.virginia.gov/VDOE/Assessment/home.shtml#Standards\\_of\\_Learning\\_Tests](http://www.doe.virginia.gov/VDOE/Assessment/home.shtml#Standards_of_Learning_Tests)
- Wainer, H. (Ed.) (2000). *Computerised Adaptive Testing: A Primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Wang, S., Jiao, H., Young, M., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, 67 (2), 219-238.
- Wang, S., Jiao, H., Young, M., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, 68 (1), 5-24.
- Web3D Consortium (2007, 2008) ISO/IEC FDIS 19775:2008, Information technology - Computer graphics and image processing - Extensible 3D (X3D); ISO/IEC 19776:2007, Information technology - Computer graphics and image processing - Extensible 3D (X3D) encodings; ISO-IEC-19777-1-X3DLanguageBindings-ECMAScript & Java.
- Webb, N. (1995). Group collaboration in assessment: Multiple objectives, processes, and outcomes. *Educational Evaluation and Policy Analysis*, 17(2), 239.
- Weiss, D. & Kingsbury, G. (2004). Application of computer adaptive testing to educational problems. *Journal of Educational Measurement*, 21, 361-375.
- Williamson, D. M., Almond, R. G., Mislevy, R. J., & Levy, R. (2006). An application of Bayesian Networks in automated scoring of computerized simulation tasks.
- Williamson, D. M., Mislevy, R. J., & Bejar I. I. (2006). (Eds.), *Automated scoring of complex tasks in computer-based testing* Mahwah, NJ: Erlbaum.
- Willighagen, E. & Howard, M. (2007). Fast and scriptable molecular graphics in web browsers without Java3D. *Nature Precedings* 14 June. doi:10.1038/npre.2007.50.1.  
<http://dx.doi.org/10.1038/npre.2007.50.1>
- Wirth, J., & Funke, J. (2005). Dynamisches Problemlösen: Entwicklung und Evaluation eines neuen Messverfahrens zum Steuern komplexer Systeme. In E. Klieme, D. Leutner & J. Wirth (Eds.), *Problemlösekompetenz von Schülerinnen und Schülern* (pp. 55-72). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Wirth, J., & Klieme, E. (2003). Computer-based assessment of problem solving competence. *Assessment in Education: Principles, Policy & Practice*, 10 (3), 329-345.
- Xi, X., Higgins, D., Zechner, K., Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater v1.0* (RR-08-62). Princeton, NJ: Educational Testing Service.

White Paper 3: Technological issues for computer-based assessment

Zhang, Y., Powers, D. E., Wright, W., & Morgan, R. (2003) *Applying the Online Scoring Network (OSN) to Advanced Placement Program (AP) tests* (RM-03-12). Princeton, NJ: Educational Testing Service. Retrieved August 9, 2009 from <http://www.ets.org/research/researcher/RR-03-12.html>